JMRA リサーチイノベーション委員会 2022 年度研究活動

マハラノビス研究会報告

2023 年 3 月 31 日

朝野熙彦

目　次

キーワード:汎距離、多重共線性、一般逆行列、判別分析、MT システム

１章　問題意識

１.１　研究の背景

マーケターは、複数の測定変数を座標に用いて、自社や競合の製品やサービスを空間にマッピングすることがある。関心のある製品やサービスがどれほどユニークなのか、あるいは集団内に埋没しているかを知ることができる。マッピングはポジショニング戦略そしてセグメンテーション戦略を立てる基礎としても使われる。

ところでマッピング情報を集約する過程には測定変数の相関が影響してくる。たとえばノート PC をマッピングするためにディスプレイサイズ、重量、CPU 速度の 3 変数を選んだとしよう。ディスプレイサイズと重量の相関は高いだろうから、情報を集約すれば、ノート PC は概ね大きさで区別されることになる。我々にとってなじみ深いユークリッド距離も差の二乗和をとるロジックなので同様に相関の影響を受ける。

従来のリサーチデータの解析では、測定変数の組から相関の高い変数を除いてバランスをとるとか、あらかじめ因子分析で空間を直交化してから因子得点をクラスター分析にかけるという処理がしばしば行われてきた。しかしながらマッピング分析の利用者としては、できれば評価項目をすべて利用したいだろう。また因子や主成分のような潜在変数よりも、測定変数をそのまま座標に使うほうが解釈が容易だろう。

そこで、測定変数間の相関情報を考慮に入れながら評価対象が遠いか近いかが測れないだろうか、という問題意識が生まれる。

１.２　研究する統計モデル

Mahalanobis（1936）は変数間の相関を考慮した汎距離を提案した[1]。マハラノビスの汎距離は、R や Python など多くのプログラムに関数として組み込まれている。さらにマハラノビスの汎距離を標準的に出力する商用ソフトも多い。

マハラノビスの汎距離は歴史的な存在のように思われてきたかもしれないが、近年になって品質工学の分野で注目されるようになった。MT（マハラノビス・タグチ）法と呼ばれる一連の技法が生産管理の現場に普及してきたのは 2010 年ごろからである。本研究会の開始にあたって我々は次のような疑問を持った。

① MT 法が生産管理で成功したのはなぜか。MT 法でいう単位空間とマーケティングのマッピング空間はどう違うのか。

②MT法に余因子展開を組み入れたのが MTA 法であるが、MTA 法によって多重共線性の問題は解決するのだろうか。

③多変量解析にマハラノビスの汎距離を導入することでどのような利点があるのだろうか。

これらの素朴な疑問を含めて汎距離の理論と応用にかかわる疑問に答えるために本研究を行った。

２章　MT システムの成功要因

２.１　田口玄一の着想

MT システムは田口玄一が提唱した品質工学の方法論の総称である。MT はマハラノビスとタグチの頭文字だが、MT システムは 2 人の共同研究ではなく、田口が独自に開発したものである。

田口は米国や日本の製造業でタグチメソッドの普及に貢献した技術志向の研究者である（田口 1999, 145 頁）。生産分野には固有の課題があり、また利用可能なデータにも制約がある。田口はアンナ・カレーニナの冒頭の一節から MT システムの着想を得たと言われている[2]。

> 幸福な家庭はすべて互いに似かよったものであり、不幸な家庭はどこもその不幸のおもむきが異なっているものである

この文意を品質工学の文脈で言い替えれば、次の 2 つの言明を意味するだろう。

A）順調に製品を生産しているときの検査品の測定データは似ている
B)その一方でトラブルの原因は多様なので、不良品は様々な点で異なる

図 2.1 に「単位空間」と「信号空間」という MT システムの概念を示す。単位空間とは、品質管理の専門家が基準とみなしたプロダクトの集合である。一方、信号空間は良品か否かが不明なプロダクトの集合をさす。単位空間で不良品の識別モデルを作り、信号空間にそれを適用する。機械学習でいえば、前者が学習フェーズ、後者が予測フェーズに該当する。

品質管理においては、仮に検査では合格として出荷しても、製品使用時にも良品かどうかは分からない。この不確実性に対して従来の品質検査では、検査項目を増やし、それぞれの許容範囲を厳格化する方向で対処してきた。MT システムは従来とは異なるアプローチで不良品検出の能力アップを目指したシステムである。

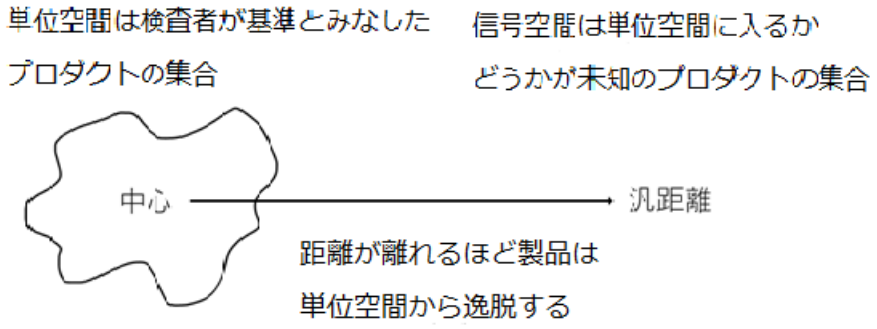


図 2.1　品質工学で扱う2つの空間

2.2　相関を考慮した距離

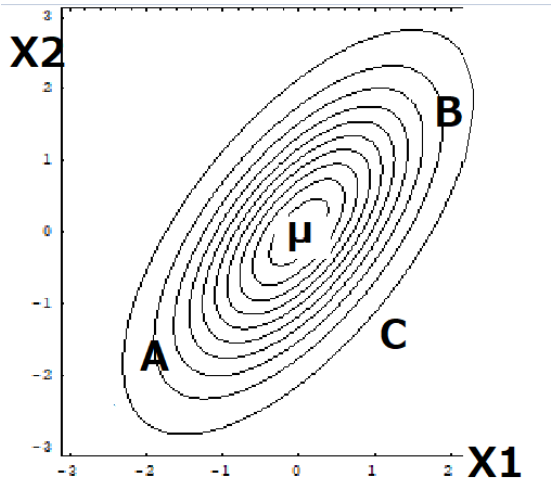


図 2.2 に MT システムの概念を示す。検査 X1 と検査 X2 に相関がある場合の検査データの同時分布を描いたものである。図の楕円は出現可能性の高さを表し、中心 μ の周辺ほど多くのプロダクトが出現すると想定したモデルである。μ は製品の目標仕様であることが期待される。一番外側の楕円が外れ値をカットオフするための閾値ラインだとしよう。

図 2.2　同時分布に着目した外れ値の検出

さて同図の検査対象 A,B,C であるが、検査 X1、X2 のどちらでみても{A＜C＜B}の順の測定値が得られている。製品 C は A と B の中間に入る。実際ユークリッド距離で測れば C は中心 μ に近い。ところが出荷後トラブルを起こす可能性があるのは製品 C である。μ を中心にした等高線に着目すれば、製品 C は閾値ラインの外に出るからである。

このように説明変数間の相関を利用した距離が MT システムでいうところのマハラノビスの汎距離であった。本当にマハラノビスが提唱したものだったかどうかは 4.1 節で検討する。

MT システムの目的は中心からの逸脱度を汎距離という単一の尺度で測定して、不良品を検出することにあった。変数間の相関を考慮することによって、検査項目を増やすことなく、従来の判定法よりも高い精度で不良品を検出できる。これが品質検査の分野で MT システムが成功した理由であった。Taguchi ら (2002)や田口・兼高 (2002)に見るように、いわゆるマハラノビスの汎距離は生産管理の分野で活躍してきた。

2.3　マーケティングからの視点

統計解析は応用場面によって使い方も解釈も違ってくる。良品か不良品かを分類することと、マーケティングで市場を分割することは似て非なる行為である。ここではマーケティングの視点から論点を列挙しておこう。

①一口に不良といっても過大と過小では修正方向が異なる。不良品を検出しただけでは解決にはならない。品質管理と違ってマーケティングには方向概念が必

要である。図 2.2 でいえばＡとＢは汎距離としては同じだが方向は対極にある。

②マーケティングの市場認識では幸福な家庭でさえ単一の集団とは限らない。資生堂と花王ソフィーナのユーザーは異なるのかもしれない。

③マッピング空間では単位空間と信号空間の両要素を同時にプロットするのが通常である。市場の現行品のかたまりを単位空間と考えれば、それから離れた製品やサービスそして消費者は、ただの逸脱ではなく有望なプレーヤーかもしれない。

④MT システムにおける単位空間は検査者が外包的に定めたものである[3]。調査の言葉でいえば検査者が該当サンプルを有意選出(purposive selection)した結果である。その点、マーケティングでは未顧客や想定外の生活者に関心をもつことがある。市場には異質性があって当然だという認識がセグメンテーション論の根底にある。

３章　マハラノビスの汎距離

３.１　多次元空間における距離への着目

マハラノビス（1930）は多次元空間に２つの母集団が存在する状況の下で母集団間の距離を測ることに関心をもった。マハラノビスは集団の平均値にｍ、変数の分散にσ^2 、平均値間の距離にD^2という記号を用いた。1930 年の論文では(3.1)の距離尺度が最も使いやすいと結論づけている。

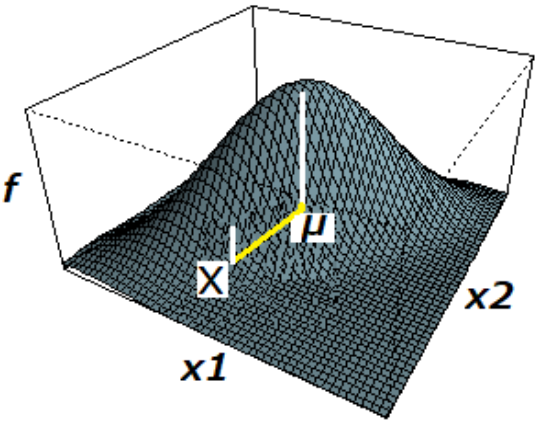
$$D^2 = \frac{1}{P}\sum_p \frac{\left(m_{p1} - m_{p2}\right)^2}{\sigma_p^2} - \frac{1}{P}\sum_p \left(\frac{1}{n_{p1}} + \frac{1}{n_{p2}}\right) \qquad (3.1)$$

(3.1)では変数の違いを小文字の p、変数の総数を大文字の P で表している。マハラノビスは統計量から真の平方距離を不偏推定するために第 2 項の補正項を導入した。2つの集団から得られたデータ数n_{p1}, n_{p2} の添字には変数番号pと母集団 1 と 2 の区別が入っている。なぜ測定変数によってデータ数が変化すると想定したかについては説明がない。

多変数の情報を集約する以上、各変数の測定スケールの違いが影響する。メートルで測っていたデータをミリに直せば桁が 1000 倍になってしまう。マハラノビスは測定データを標準偏差で割ることで各変数のスケールを調整した。それが(3.1)の第 1 項の規準化の目的であった。

平均値の差の二乗和が大きいほど(3.1)のD^2は大きな値をとる。しかし(3.1)では、変数間の共分散を考慮していない。したがって 1930 年の提案は後年の汎距離とは異なる。しかし 2 つの母集団の平均値間の距離を測定したいという問題意識は共通している。[4]

3.2　多変量正規分布の仮定



3.3 節以降のために多変量正規分布の密度関数をここで示しておく。

変数が $x_1, x_2, \cdots, x_p$ とp個ある場合に、p次元正規分布の密度関数 f を(3.2)で表す。

図 3.1　多変量正規分布の密度関数

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \qquad (3.2)$$

行列とベクトルはボールド体で書く。p×pの分散共分散行列 $\boldsymbol{\Sigma}$ の逆行列を $\boldsymbol{\Sigma}^{-1}$ と書き、$\boldsymbol{\Sigma}$ の行列式を $|\boldsymbol{\Sigma}|$ と書く。(3.2)の π は円周率である。(3.2)のプライム′ は行列あるいはベクトルを転置する操作を表す。

(3.2)の確率変数 $\boldsymbol{x}$ がp次元正規分布に従うとすれば多変量正規分布はp次の期待値ベクトル $\boldsymbol{\mu}$ とp×p の分散共分散行列 $\boldsymbol{\Sigma}$ で確定する。このことを

$\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$　で表す。従って(3.2)は期待値からの差のベクトル $(\boldsymbol{x}-\boldsymbol{\mu})$ をスカラー f に写像する関数ともいえる。 図 3.1 は $p=2$ の場合について $\boldsymbol{x}$ と $\boldsymbol{\mu}$ に対応した密度関数の値を縦棒で描いたものである。指数関数 exp の{　}内の

$(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ は $\boldsymbol{\Sigma}^{-1}$ が対称行列の場合に2次形式と呼ばれる。$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ は理論モデルにおける母数なので標本データから推定する場合は、それぞれを標本平均ベクトル $\boldsymbol{m}$ と標本分散共分散行列 $\boldsymbol{S}$ に置き換える。

3.3　マハラノビスの原案

マハラノビス(1936)は２つの母集団の分布が(3.3)の多変量正規分布に従うと仮定した。

$$\boldsymbol{x}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{x}_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \qquad (3.3)$$

マハラノビスは２つの多変量正規分布は平均ベクトルだけが異なり、分散共分散行列は等しいと仮定した。その上で２つの母集団の平均値間の距離を次のように定義して、これをΔ^2 統計量と名付けた。

$$\Delta^2 = \frac{1}{p}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \qquad (3.4)$$

$\Delta^2 \geq 0$ である。また各変数の標準偏差が 1 で、しかも無相関の場合は $\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}$ であるから、(3.4) 右辺の 2 次形式は平方ユークリッド距離 $d^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ に一致する。Δ^2 がユークリッド距離を特殊な場合として含むことからマハラノビスは(3.4)を汎距離と呼んだ。

原著では (3.4) をさらに余因子行列に展開しているが、余因子展開に意味があるかどうかは 5.1 節で検討する。

さて、(3.4)は理論モデルの式であって、実際に標本データから計算する場合は(3.5)の標本統計量D^2によってΔ^2を推定するというのがマハラノビスの原案であった。

$$D^2 = \frac{1}{p}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \frac{2}{\bar{n}} \qquad (3.5)$$

ここで$\bar{n}$ は 2 群のサンプル数n_1, n_2 の平均値である。$\bar{n}$ が大きくなれば (3.5)の第 2 項はゼロに近づくのでD^2はΔ^2 に近づく。

なおマハラノビスは$\boldsymbol{\Sigma}$が 2 群間で一定という制約をおかなければ汎距離が一般化できるという見通しを述べている。

マハラノビス(1949)自身がD^2統計量の研究をレビューしている。同じ 1930 年代にはフィッシャーによる生物統計学の判別分析があったことが紹介されているが(3.5)の一般化については言及がない。マハラノビスの原案で気になった点は次の2つである。

■マハラノビス自身が今後の課題を述べたように、群によって$\boldsymbol{\Sigma}$が異なる場合に汎距離を一般化する意義はあるだろう。この点は 4.4 節で検討しよう。

■マハラノビス自身の定義も変遷している。(3.1)にあったn_{p1}, n_{p2}は(3.5)では変数の違いを指す添字が消えた。また$1/p$と$2/\bar{n}$の係数は、その後の統計学の論文では削除されていることが多い。次章であらためてD^2の定義を検討しよう。

４章　汎距離の各種の定義

４.１　汎距離の全体像

マハラノビスの提唱を受けて、その後多くの研究者がマハラノビスの汎距離を分析モデルに組み込み実証研究の道具として利用してきた。

しかし一口に「相関を考慮したＡとＢの距離」と言ってもそのＡとＢが何をさすかによって、そして母集団について何を仮定するかによって汎距離の定義は変化する。

もちろんマハラノビスの原案を拡張発展させることは研究として自然なことである。しかし同じ名称を使いながら研究者によって定義が異なることは混乱や誤解を招くに違いない。

まず各種の汎距離の関係を図 4.1 に示す。ここでは多変量正規分布の母数である期待値を母平均と呼んだ。期待値を測定データから推定する場合は群平均と呼んで区別した。同様に母数である分散共分散行列は$\boldsymbol{\Sigma}$、それを測定データから推定する場合は$\boldsymbol{S}$と書いて区別した。

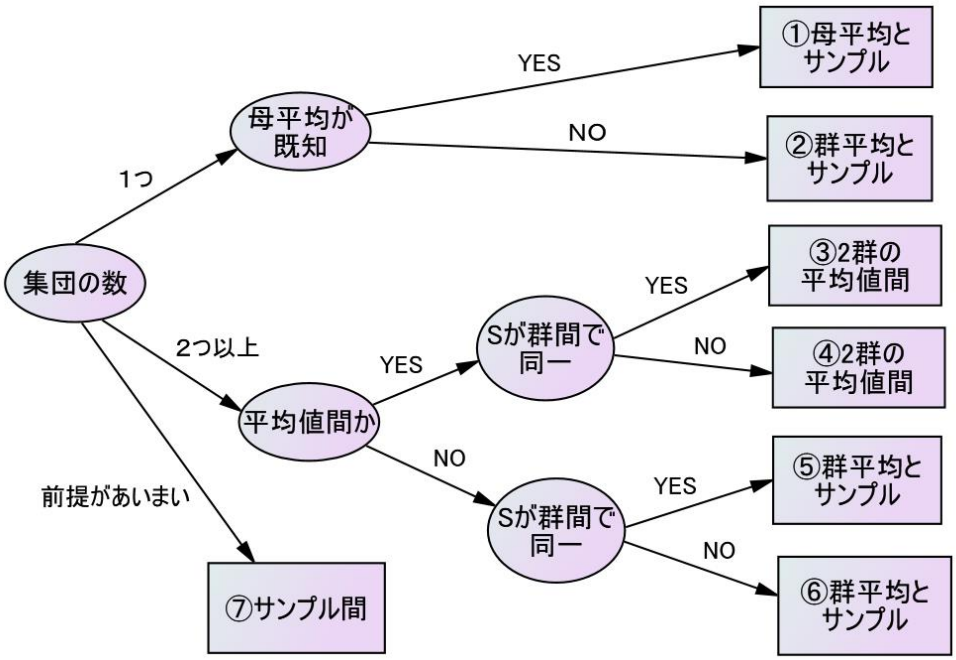


図 4.1　各種の汎距離の間の関係

図 4.1 で①~⑦と区別した 7 つの汎距離の定義式を次に記述する。

【ケース１】１つの母集団の母平均からサンプルまでの距離

$(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$　多変量正規分布の母平均と分散共分散行列が既知の場合に、$\boldsymbol{\mu}$ から確率変数の実現値 X＝$\boldsymbol{x}$ までの距離を測る場合。理論モデルの下でシミュレーション分析する場合もケース１に含めることにした。

【ケース２】１つの群平均からサンプルまでの距離

$(\boldsymbol{x}-\boldsymbol{m})' S^{-1}(\boldsymbol{x}-\boldsymbol{m})$　１つの群平均$\boldsymbol{m}$ から個々のサンプルまでがどれだけ離れているかを測る場合である。MT システムでは逸脱品を発見するためにケース 2 の汎距離を用いた。現実の世界ではパラメータは未知なので、平均と分散共分散行列をデータから推定するしかない。

なお平均とは呼ばず重心や中心あるいはセントロイドと呼び、群ではなく集団やクラス、セグメントなどと呼ぶことがある。応用する分野によってそれぞれ慣習化しているので、どれが正しくどれが誤りとはいえない。

【ケース３】2 群の平均間の距離

マハラノビスが 1936 年に提唱したのが 2 群の平均間の距離だった。

$\Delta^2 = \frac{1}{p}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$　ただしマハラノビス以後の統計学の論文や書籍では通常$1/p$ の係数を除いている。データ解析においてはΔ^2 ではなく標本統計量　$D^2 = (\boldsymbol{m}_1 - \boldsymbol{m}_2)' \boldsymbol{S}^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$　を用いる。マハラノビスの原案に従って$\boldsymbol{S}$は２群間で等しいと仮定する。

【ケース４】群によってSが異なる２群の平均間の距離

S が群によって異なる場合である。本稿では、所属群で条件づけられた次の統計量を提案する。群番号を$h \neq g$ としてg 群の平均からh 群の平均を測る場合は

$$D^2(\boldsymbol{m}_h, \boldsymbol{m}_g | G = g) = (\boldsymbol{m}_h - \boldsymbol{m}_g)' S_g^{-1}(\boldsymbol{m}_h - \boldsymbol{m}_g)$$

限られたサーベイの範囲であるが、このケースに該当する先行研究と応用事例は見つかっていない。4.4 節でケース 4 の距離の性質を検討する。

【ケース 5】多群の群平均とサンプルとの距離

$$(\boldsymbol{x} - \boldsymbol{m}_g)' \boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_g) \qquad \mathrm{g} = 1, 2, \cdots, G \qquad G \geq 2$$

複数の群が存在するがそれらの分散共分散行列は共通とする。各群の平均から個々のサンプルまでの距離をこの定義式で測り、一番距離が小さい群にサンプルを判定する目的で利用することが多い。

【ケース 6】平均と分散共分散行列が群によって異なる場合の距離

$(\boldsymbol{x} - \boldsymbol{m}_g)' S_g^{-1}(\boldsymbol{x} - \boldsymbol{m}_g)$　ケース 5 の制約をゆるめた距離になる。平均からの距離が一番近い群にサンプルを判定する方式は同じである。

【ケース 7】空間の任意の２サンプル間の距離

$(\boldsymbol{x}_i - \boldsymbol{x}_j)' S^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)$　いくつかの文献で多次元空間の任意の 2 点間で距離が測定できるという記述がある。4.5 節でこのケースについて検討する。

表 4.1　いくつかの文献にみられる汎距離

		1	2	3	4	5	6	7	
発表年	著者・タイトル	1つの母平均とサンプル	1つの群平均とサンプル	2群の平均間の距離	2群の平均間の距離によってSが異なる	多群の群平均とサンプル	多群の群平均とサンプルによってSが異なる	空間の任意の2サンプル	定数による除算があるか
1936	Mahalanobis,"generalized distance"			◎					p
1971	奥野・他「多変量解析法」	●		◎		●			
1972	竹内・柳井「多変量解析の基礎」		●	◎		●		●	
1983	田中・脇本「多変量統計解析法」	●				●	●		
1986	柳井・他「多変量解析ハンドブック」	●		◎					
1994	柳井「多変量データ解析法」			◎					
1996	Koschnick, "Dictionary of Social and Market Research"			◎					
1996	水野「多変量データ解析講義」			◎				●	
1999	田口「品質工学の数理」		●						k
2000	Wedel,Kamakura,"Market Segmentation"							●	
2000	Taguti et.al."The Mahalanobis-Taguchi System"		●						k
2001	永田・棟近「多変量解析入門」		●	◎		●	●		
2002	Venables,Ripley,"Modern Applied Statistics with S"						●		
2005	Cerioli,"K-means cluster analysis and Mahalanobis metrics"		●				●	●	
2011	豊田・池原、変数間の関係性を考慮してクラスター数を決定するk-means法の改良						●		
2012	Nelson,"On K-means clustering using Mahlanobis distance"						●	●	
2021	Yonenaga,Suzukawa,"Bayesian estimation for misclassification rate in linear discriminant analysis"			◎					
2022	宮川・永田「タグチメソッドの探求」	●				●	●		

◎はマハラノビスが提唱した汎距離

●はマハラノビスが提唱していない汎距離

表 4.1 にいくつかの文献の中で、どの汎距離が扱われてきたかを一覧している。マハラノビスの距離という名称でマハラノビスの原案だけを紹介した文献は表 4.1 では柳井(1994)と Koschnick(1996)、Yonenaga ら(2021) だけである。奥野・他 (1971) の「多変量解析法」は多変量解析に関する古典的な名著だが、ケース 1 (266 頁)、ケース 3 (290 頁)、ケース 5 (302 頁) をすべてマハラノビスの汎距離と呼んでいる。このように 1 冊の本の中で複数の定義が出てきたら読者は混乱しないだろうか。

さて MT システムにおいて田口 (1999) 自身がマハラノビスの汎距離と呼んだのは表 4.1 のケース 2 であった。MT システムの提唱にあたっては、MT システムでいう汎距離がマハラノビスの原案とは異なることを明記する方が良かったであろう。

表 4.1 の右端の欄は 2 次形式を変数の数で割ったか否かを示したものである。変数の数を p と書くか k と書くかは本質的な問題ではない。表 4.1 の多くの文献で変数の数による除算がない。また R の stats ライブラリの mahalanobis 関数も、Python の scipy ライブラリの distance.mahalanobis 関数も変数の数で割るという処理はしていない。

理論モデルの世界では $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ であれば $\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ は自由度pのカイ二乗分布に従う。しかし宮川・永田 (2022、123 頁) は、現実のデータ解析においては推定値 D^2 はカイ二乗分布に従わないと指摘している。そこで我々も 2 次形式を p で割らないことにする。厳密には平方汎距離 squared generalized distance と呼ぶべきだが、「平方」であることはいちいち断らない。

本節の最後に表 4.1 と図 4.1 についてさらに補足する。

■ 母集団の意味

汎距離のケース 1 と 2 の違いは、データ解析の実務でいえば分析者が分布のパラメータを自分で指定する場合がケース 1 で、測定データから平均と分散共分散行列を推定する場合がケース 2 になる。

数理統計学では通常、母集団と確率分布を同一視するが、社会調査では有権者や地域住民の集団を母集団と呼んでいる。そしてマーケティングでは消費者集団を母集団と扱うことが多い。同じ母集団といっても、分野によって概念が異なることに注意したい。一方 MT システムには単位空間や信号空間という集合が出てくるが、これも確率分布を前提としたものではない。

なおケース 1 と 2 を場合分けするのであれば、ケース 3～6 も母集団のパラメータが既知か否かで場合分けすべきかもしれない。その場合はケースの数がさらに 4 つ増えることになる。

■ 群に共通した分散共分散行列 S の推定法

ケース 3 と 5 では各群の分散共分散行列が同一であることを仮定している。

現実の標本データにおいて各群の S が厳密に一致することは考えられないので pooled covariance matrix S を標本データから計算することが行われている。具体的にはグループ g 別に求めた S_g を各群のサンプル数で加重平均して S を求める。2 群の場合の計算法は田中・脇本(1983、114-115 頁)では(4.1)のように示されている。

$$S = \frac{1}{n_1 + n_2 - 2}\left\{(n_1 - 1)S_1 + (n_2 - 1)S_2\right\} \tag{4.1}$$

その類推として 3 群の場合なら次式になるだろう。

$$S = \frac{1}{n_1 + n_2 + n_3 - 3}\left\{(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3\right\}$$

ところで市場に複数の群が存在するにもかかわらず、複数であることを無視してデータ全体で S を計算するとどうなるだろうか。部分と全体では異なる結論が導かれることを警告したのがシンプソンズパラドックスであった。図 4.2 の A は 2 群とも測定変数間に負の相関がある場合、B は 2 群で相関が正負逆になる場合である。上記のプーリング式を用いれば A)では負の相関、B）では無相関になる。

しかし群の違いを無視して全データを一括して計算すると A,B とも正の相関になる。結論として散布図が A のケースでは（4.1）で S_1, S_2 をプーリングして S を求めることに意味がある。一方 B のケースでプーリング法を使うことは不適切である。

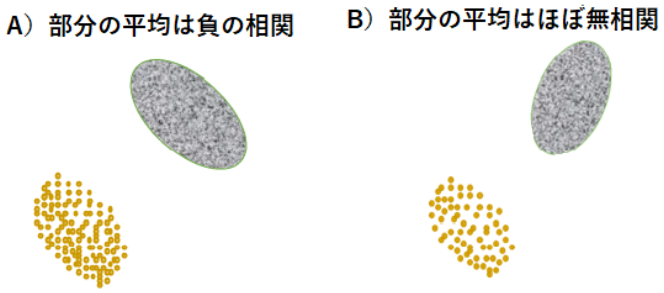


図 4.2　全データを一括すれば正の相関

4.2　１つの母集団の中心からの汎距離

汎距離の中でも解釈しやすいのがケース１と２の汎距離である。１つの多変量正規分布を図 4.3 のような山形でイメージしよう。山頂から出発してどれだけ高度を下げたかという高度差で評価したのが汎距離である。下山の場面を考えると、ふもとに降りるほど高度が下がるので、それだけ山から離れた証拠に

なる。逆に尾根伝いに移動すれば、ほとんど高度は下がらない。

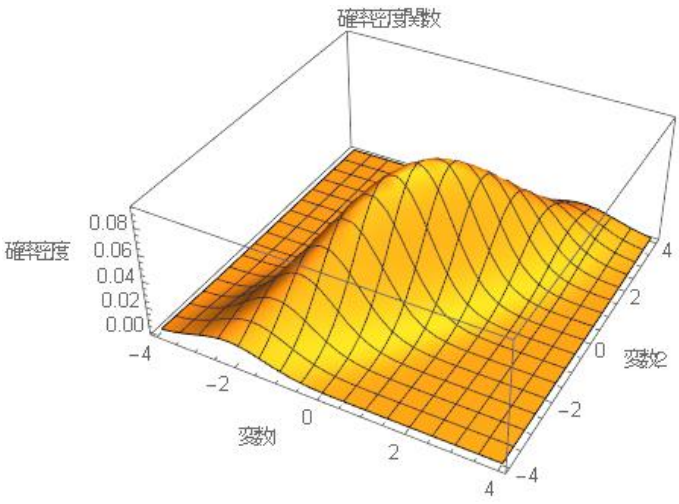


図 4.3　単峰の多変量正規分布

次に傾斜の違いが分かりやすいように p=1 の場合に距離 $D=|x|$ の密度関数を示したのが図 4.4 である。正規分布については $x \geq 0$ の領域だけを描いた。汎距離 D は非負なのでその確率分布は「半正規分布」と呼ばれる。半正規分布の密度関数は $f(D)=\sqrt{\frac{2}{\pi}}\exp\left(-\frac{1}{2}D^2\right)$ であり標準正規分布 N(0,1)の 2 倍の大きさになる。

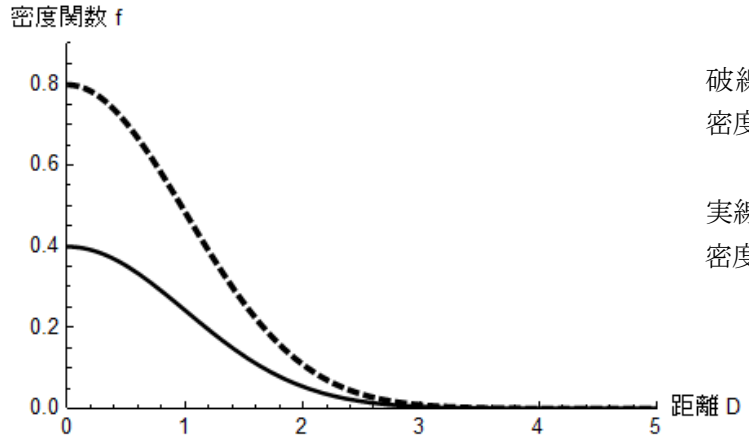


図 4.4　汎距離および正規分布の密度関数

ケース 2 の汎距離を計算する R のコードを次の囲みに示す。原データを平均偏差化することも、規準化することも必要ない。もし平方でない汎距離が必要ならば D2 の平方根をとればよい。

```
# n 行 p 列の原データ行列 X から出発して汎距離を計算する
mu <- colMeans(X)     #平均ベクトル
S <- cov(X)               #分散共分散行列
# 集団の平均からの平方汎距離
D2 <- vector("double", n)     # 計算結果を格納するベクトル
for (i in seq_len(n)) {
  # 平均からの差のベクトルと S の逆行列を使って平方汎距離を計算する
  D2[[i]] <- (X[i,]-mu) %*% solve(S) %*% (X[i,]-mu)
}
#   データフレーム X_df に汎距離 D2 を出力
X_df <- as.data.frame(X)
X_df$D2 <- D2          # 計算結果を格納
head(X_df,10)
```

4.3　複数の母集団の中心からの距離を測る Python のコード

複数の母集団に対応した汎距離がケース 5 と 6 である。ケース 5 では各群の分散共分散行列が等しいと仮定するが、ケース 6 にはこの仮定がない。

たとえば G=3 の場合で説明すると、g1 の中心、g2 の中心、g3 の中心から個々のサンプル i までの距離を測り、距離が近い集団にそのサンプルを判別することができる。

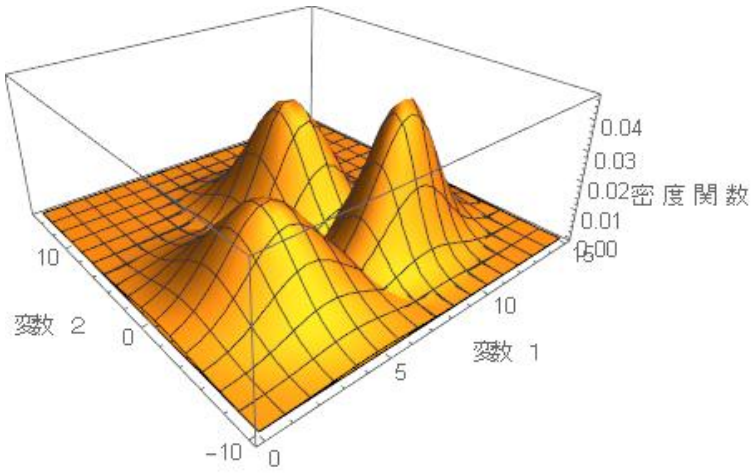


図 4.5　3 群の分布の鳥瞰図

松波成行は python を用いて以下の実験を行った。実験の目的は次の疑問に答えることにあった。

①Scipy ライブラリの distance.mahalanobis メソッドから得られる距離はケース 6 の定義式の距離と一致するか

②判別分析に適用すれば、汎距離はユークリッド距離よりも判別精度が高まるか

具体的な解析データにはフィッシャー（1936）のアヤメという統計学でよく知られたデータを用いた。setosa, versicolor, virginica という 3 種類のアヤメについてガクの長さと幅の 2 変数で測定したデータを分析した。

setosa の平均からの汎距離を計測する Python のコードを以下に示す。

$D^2\left(\boldsymbol{x},\boldsymbol{m}_g\right)=\left(\boldsymbol{x}-\boldsymbol{m}_g\right)' S_g^{-1}\left(\boldsymbol{x}-\boldsymbol{m}_g\right)$ と distance.mahalanobis の出力が一致することが囲みのコードで確認できる。distance.mahalanobis の値は D^2 ではなく D であった。

```
import numpy as np
import pandas as pd
from scipy.spatial import distance
#データセットを読み込み setosa のガク長とガク幅だけのデータセット
#setosa を作る
df=pd.read_csv("iris.csv")
setosa = df.iloc[:50,[0,1]].to_numpy()
# setosa のガク長とガク幅に関する群平均
mu = np.mean(setosa, axis=0)
# setosa のガク長とガク幅の分散共分散行列
cov_matrix = np.cov(setosa.T)
# 分散共分散行列の逆行列を numpy の linalg.pinv メソッドで求める
cov_i = np.linalg.pinv(cov_matrix)
#  ガク長とガク幅が(5.8cm, 3.5cm)のテストデータとの汎距離を測る
test = np.array([5.8,3.5])
# テストデータと群平均値との差のベクトル
delta = test - mu
# numpy ライブラリの dot メソッドを用いて平方汎距離を計算する
D2 = np.dot(np.dot(delta, cov_i), delta)
D=np.sqrt(D2)
print(D)
#scipy ライブラリの distance.mahalanobis モジュールから出力される汎距離
md = distance.mahalanobis(test, mu, cov_i)
print(md)
```

■　データからテストデータランダムに 10%抽出し、残りのデータで平均と分散共分散行列を推定する。テストデータについて分類の予測値と真の分類をクロス集計した結果を表 4.2 に示す。setosa を汎距離で予測した場合は誤分類はないが、ユークリッド距離では誤分類が 2 件発生した。クロス集計表の主対角要素の和の全体に対する比率を正解率とすれば、汎距離の正解率は 73.3%でユークリッド距離は 66.7%なので汎距離の方が優れていることになる。もちろんこの結果はテストデータに依存することに留意しなければならない。

表 4.2　判別精度の比較

(A)汎距離を使った場合

		真の分類		
		setosa	versicolor	virginica
予測分類	setosa	5	0	0
	versicolor	0	4	3
	virginica	0	1	2

(B)ユークリッド距離を使った場合

		真の分類		
		setosa	versicolor	virginica
予測分類	setosa	5	1	1
	versicolor	0	3	2
	virginica	0	1	2

4.4　複数の母集団の中心間の汎距離

マハラノビスが提案したのは表 4.1 でいえばケース 3 の「2 群の平均値間の距離」だった。では群が 3 つ以上あった場合はどうだろうか。群が$G \geq 3$ の場合に G から 2 をとる組み合わせについて中心間の距離を測ることはできる。どの群とどの群が近いかを知ることは、たんに市場全体が G 個の群に分割されていた、というよりも市場理解が深まることは明らかだろう。

ここで問題になるのは各群の分散共分散行列が等しくないケース 4 の場合である。具体例でみてみよう。図 4.6 は 3 種類のアヤメの群平均値（★印）を図示したものである。がくの長さは平均して setosa、versicolor、 virginica の順に長くなっている。図中の楕円は汎距離が$D=4$の等距離線を図示している。多変量正規分布の仮定の下では、密度（densities）が等しい線であるから、等高線といってもよい。

アヤメのデータでは分散共分散行列が群間で異なるが、ではどのように平均間の距離を測ればよいのだろうか。ここでは一方の群の平均から、他方の群の平均の位置を一つのサンプルとみなして距離を測定することにする。つまり群番号を$h \neq g$ として所属群で条件づけた次の距離を定義する。

$$D^2\left(\boldsymbol{m}_h, \boldsymbol{m}_g \middle| G=g\right) = \left(\boldsymbol{m}_h - \boldsymbol{m}_g\right)' S_g^{-1} \left(\boldsymbol{m}_h - \boldsymbol{m}_g\right)$$

これがケース 4 の汎距離である。この距離には図 4.6 に見るように対称性がない。例えば、setosa の平均値からみると versicolor の平均値は$D=4$ のラインを超えるが、逆に versicolor から見れば、setosa の平均値は$D=4$ のラインに入る。

ユークリッド距離の場合は A と B の距離と B と A の距離の間には

$d\left(A,B\right)=d\left(B,A\right)$ という対称性が成り立つ。しかしケース 4 の汎距離は一般に$D^2\left(\boldsymbol{m}_h, \boldsymbol{m}_g \middle| G=g\right) \neq D^2\left(\boldsymbol{m}_g, \boldsymbol{m}_h \middle| G=h\right)$　であり距離の公理を満たさない

[5]。このような非対称性はマーケティングではサプライチェーンのプレーヤー間でも、あるいは企業と顧客の関係性についても生じ得る。俗にいう片想いの現象が起きる理由をグラフィックに表すことができる。

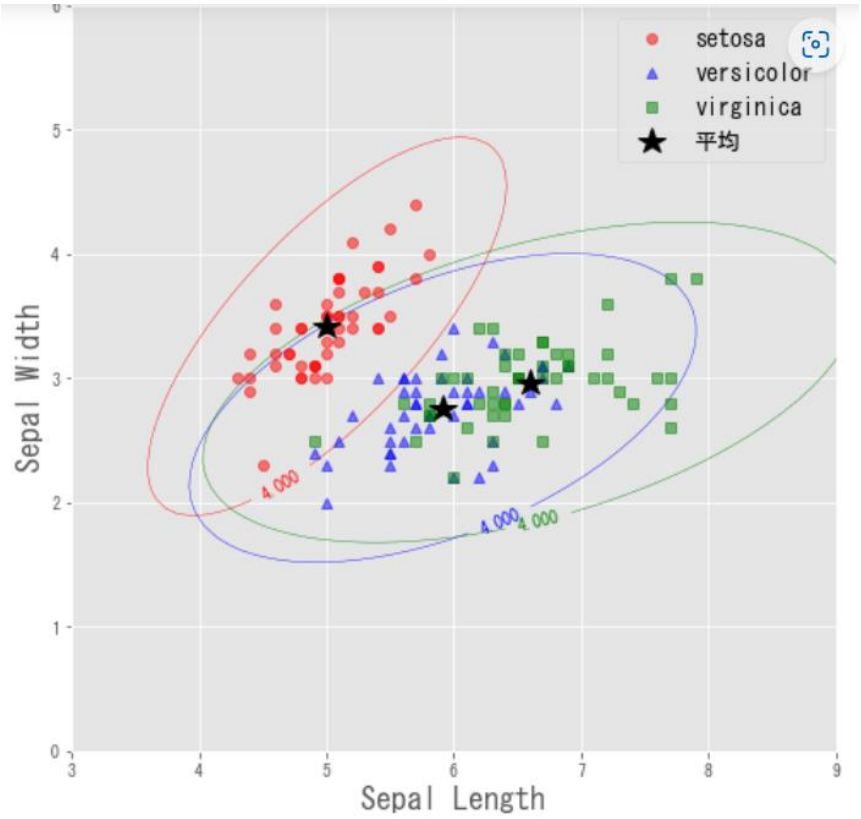


図 4.6　3 種類のアヤメの等距離線

4.5　任意の２点間の距離

我々はｐ次元空間の任意の２点間の汎距離をケース 7 の汎距離と呼んだ。水野（1996、216 頁）の記述を見ると、2 点間の汎距離とは複数の母集団が存在する場合は個々の集団内で分散共分散行列を計算して、それぞれの集団内で点間距離を測る指標だと読める。水野の文脈では複数の母集団とはクラスターを指している。

$$D^2\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)'\boldsymbol{S}^{-1}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right) \qquad (4.2)$$

ケース 1~6 の汎距離は常に集団の中心からの距離を測ってきたので、多変量正規分布の仮定を認めるならば、汎距離は「密度の変化分」を表した。しかしケース 7 の汎距離は中心からの距離ではない。では空間の任意の 2 点間の距離を測る理論的な根拠はあるのだろうか。水野は（4.2）が導かれる論拠を示していない。

この疑問に対して竹内・柳井（1972、pp.278-279）は次のように論拠を示している。まずN次元の単位ベクトルを考える。単位ベクトルとはi番目の成分だけが1で残りが0のN次のベクトル$\boldsymbol{e}_i\left(i=1,2,\cdots,N\right)$である。その役割はi番目のサンプルをその他のサンプルと識別することにある。

次に任意の2サンプルの差のベクトル$\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)$を$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$で張られる部分空間に射影する。ただし各$\boldsymbol{x}$は平均偏差化されていると仮定する。

$\boldsymbol{X}=\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\right)$として射影子$\Pi_X$を用いて射影ベクトルのノルムの2乗のN倍を求めると

$$
\begin{aligned}
&N\left\|\Pi_X\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\right\|^2=N\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\\
&=\left(\tilde{\boldsymbol{x}}_i-\tilde{\boldsymbol{x}}_j\right)'N\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\tilde{\boldsymbol{x}}_i-\tilde{\boldsymbol{x}}_j\right)=\left(\tilde{\boldsymbol{x}}_i-\tilde{\boldsymbol{x}}_j\right)'\boldsymbol{S}^{-1}\left(\tilde{\boldsymbol{x}}_i-\tilde{\boldsymbol{x}}_j\right)
\end{aligned}
\qquad (4.3)
$$

ここで$\tilde{\boldsymbol{x}}_i$, $\tilde{\boldsymbol{x}}_j$は行列$\boldsymbol{X}$の第i、第j行を内容とするp次の列ベクトルである。したがって(4.3)は(4.2)と等しく、点iとjの汎距離は射影ベクトルのノルムの平方に比例する。なお部分空間への射影が成り立つには、単に線形空間であることを条件にするだけでよく特定の確率分布は仮定しない。

難しいのは(4.3)の統計学的な意味づけである。$\boldsymbol{S}$は分散共分散行列であり、$\boldsymbol{S}^{-1}$は共分散の影響を除去する役割を果たすものと解釈するのだが、そう解釈するためにはN個のデータが分散共分散行列が$\boldsymbol{S}$である同一の集団に属するものと認めなければならない。

では集団全体が複数の集団に分割されている場合はどう距離を測ればよいのだろうか。多母集団の場合に（4.3）をどう拡張するかという問題である。たとえば日本で任意の2人の汎距離を測ることができ、アメリカでも任意の2人の汎距離を測ることができたとしよう。では、日米の垣根を超えて任意の2人の間の汎距離をどう測ればよいか、という問題と同じことである。

5章　数値解析の諸技法

5.1　MTA法と余因子行列

変数の間に1次従属の関係があるときは分散共分散行列の逆行列が求められないので汎距離が計算できない[6]。田口（2002）はこの欠点を克服するために余因子行列を利用したMTA(Mahalanobis Taguchi Adjoint）法を提案した。

このMTA法は実務ガイドにも紹介されている。たとえば立林（2008、35p

頁）には 1 次従属が 1 つだけであれば逆行列問題は解決できるとしている。本当だろうかというのが本節の疑問点である。まず余因子行列を定義しよう。

$A=\left(a_{ij}\right)$ を p× p 行列とし、この行列から i 行と j 列を除いた

$(p-1)\times(p-1)$ 次の部分行列 A_{ij} の行列式を $\left|A_{ij}\right|$ と書く。行列の i 行 j 列要素に符号付きのスカラー $(-1)^{i+j}\left|A_{ij}\right|$ を配置した p 次の行列をＡの余因子行列(cofactor matrix)と呼ぶ。Harville (1997)によれば余因子行列を転置した行列はＡの随伴行列 (adjugate matrix)と呼ばれる。随伴行列を求める関数を本節では adj( )と名付けた。随伴行列については次の性質が証明されている。

$$A\cdot adj(A)=adj(A)\cdot A=|A|I_p、\qquad A^{-1}=adj(A)/|A|$$

$|A|\neq 0$ つまり A が正則の場合は随伴行列を $|A|$ で割ることでＡの逆行列が求められる。なお MTA 法ではこの随伴行列を余因子行列と呼んでいるが、分散共分散行列は対称行列なので余因子行列を転置してもしなくても数値は同じなので実害はない。

さて随伴行列を使えば分散共分散行列が 1 次従属で階数（ランク）が落ちている場合でも汎距離が計算できるのだろうか。以下 3 点の疑問を検討しよう。

（疑問１）余因子行列を使えばランクが１つ落ちた行列でも逆行列が計算できるか

変数を x_1, x_2, x_3 としてランクが１つ落ちたケースを A,B ２つ検討する。

【ケース A】　$0x_1+x_2-x_3=0$　という一次従属のケース

サンプル数 n=5 の平均偏差データ行列 X と分散共分散行列 S の数値例を示す。

$$\boldsymbol{X}=\begin{bmatrix}2&1&1\\2&3&3\\0&-1&-1\\-1&-1&-1\\-3&-2&-2\end{bmatrix},\qquad \boldsymbol{S}=\frac{1}{n}X'X=\begin{bmatrix}3.6&3&3\\3&3.2&3.2\\3&3.2&3.2\end{bmatrix}$$

S のランクは 2 で行列式は 0 である。S の随伴行列（$\boldsymbol{A}^*$）は次の通り計算できる。

$$\boldsymbol{A}^{*} = adj\left(\boldsymbol{S}\right) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2.52 & -2.52 \\ 0 & -2.52 & 2.52 \end{bmatrix}$$

MTA 法の主張が正しければ $\frac{1}{|\boldsymbol{S}|}\boldsymbol{A}^{*}$ が S の逆行列になるのだが、そもそも行列式が 0 なので $\frac{1}{|\boldsymbol{S}|}\boldsymbol{A}^{*}$ は計算できない。$|\boldsymbol{S}|$ を無視したとしても S と $\boldsymbol{A}^{*}$ の積は

$$\boldsymbol{S}\boldsymbol{A}^{*} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$ なので、これを定数倍しても対角行列にはならない。$\boldsymbol{S}^{-1}$ が導けない以上 D^{2} も求められない。それにもかかわらず、実務の現場では逆行列が計算できたという発見が現れる理由をケース B で示す。

【ケース B】 $x_1 + x_2 - x_3 = 0$ という一次従属のケース

$$X = \begin{bmatrix} 2 & 1 & 3 \\ 2 & 3 & 5 \\ 0 & -1 & -1 \\ -1 & -1 & -2 \\ -3 & -2 & -5 \end{bmatrix}, \quad \boldsymbol{S} = \frac{1}{n}X'X = \begin{bmatrix} 3.6 & 3 & 6.6 \\ 3 & 3.2 & 6.2 \\ 6.6 & 6.2 & 12.8 \end{bmatrix}$$

このケースも S のランクは 2 なので行列式は 0 である。しかし、たとえば R を使うと $|\boldsymbol{S}| = 5.595524e-15$ と出力されてしまう。理論上はゼロであるべきだが小数点以下 15 桁目に数値が出る。これは計算処理の誤差にすぎないのだが、そこで誤解が起きる。S の随伴行列 $\boldsymbol{A}^{*}$ は

$$\boldsymbol{A}^{*} = \begin{bmatrix} 2.52 & 2.52 & -2.52 \\ 2.52 & 2.52 & -2.52 \\ -2.52 & -2.52 & 2.52 \end{bmatrix}$$

プログラム上は $\boldsymbol{A}^{*}$ を Adj と書き、round(X %*% Adj, digits=10) として小数点以下 10 桁でまるめ処理すると行列の積は

$$\boldsymbol{SA}^{*}=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}$$ この行列を定数倍しても単位行列にはならないの

はケースAと同じことである。結局、ランクが落ちれば逆行列は得られない。なおケース B のデータから得られる相関係数は下記の通り、どの変数の組み合わせをみても相関係数は1ではない。このことは相関係数が1の組み合わせがないことをもって多重共線性がない、と判断することが正しくないことを意味する。

$$\boldsymbol{R}=\begin{bmatrix}1&0.884&0.972\\0.884&1&0.969\\0.972&0.969&1\end{bmatrix}$$

（疑問２）変数に定数データが含まれていた場合にD^2 は求められるのか

さきほどのケースAの変数x_3 が定数だったとしよう。平均偏差をとれば0のデータになる。平均偏差行列Xと分散共分散行列Sは次の通り。

$$X=\begin{bmatrix}2&1&0\\2&3&0\\0&-1&0\\-1&-1&0\\-3&-2&0\end{bmatrix},\qquad \boldsymbol{S}=\frac{1}{n}X'X=\begin{bmatrix}3.6&3&0\\3&3.2&0\\0&0&0\end{bmatrix}$$

ここでもSの階数（ランク）は2で行列式は0である。Sの随伴行列は次の通り計算できる。

$$\boldsymbol{A}^{*}=\begin{bmatrix}0&0&0\\0&0&0\\0&0&2.52\end{bmatrix}$$

$\left|\boldsymbol{S}\right|=0$ なのでSの逆行列は計算できない。しかも２つの行列の積

$$\boldsymbol{SA}^{*}=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}$$ を定数倍しても対角行列にはならない。結論とし

ては随伴行列が計算できたからといって$\boldsymbol{S}^{-1}$ が求められるわけではない。この小実験は、変動のない変数を予め除いてから分析を行うという、従来のデータ

解析がしてきた前処理が適切であったことを確認するものである。

（疑問３）多重共線性に近い準多重共線性の事態なら随伴行列が役立つのか

準多重共線性とは何かは 6.2 節で改めて述べる。疑問１のケース A の x_3 の一部を変更して調べてみよう。平均偏差行列 X と分散共分散行列 S が次の通りだとする。変数 x_2 と x_3 の相関係数は 0.943 であった。

$$X=\begin{bmatrix}2&1&1\\2&3&3\\0&-1&-2\\-1&-1&0\\-3&-2&-2\end{bmatrix},\qquad \boldsymbol{S}=\frac{1}{n}X'X=\begin{bmatrix}3.6&3&2.8\\3&3.2&3.2\\2.8&3.2&3.6\end{bmatrix}\qquad (5.1)$$

S の階数（ランク）は 3 で行列式は 0.88 である。S の随伴行列は次の通り。

$$\boldsymbol{A}^*=\begin{bmatrix}1.28&-1.84&0.64\\-1.84&5.12&-3.12\\0.64&-3.12&2.52\end{bmatrix}$$

随伴行列にもとづく S の逆行列は

$$\boldsymbol{S}^{-1}=\frac{1}{|\boldsymbol{S}|}\boldsymbol{A}^*=\begin{bmatrix}1.455&-2.091&0.727\\-2.091&5.818&-3.545\\0.727&-3.545&2.864\end{bmatrix}\qquad (5.2)$$

$\boldsymbol{S}\dfrac{1}{|\boldsymbol{S}|}\boldsymbol{A}^*=I$ になるから(5.2)は S の逆行列であることが確かめられる。

$D_i^2=\boldsymbol{x}_i'\boldsymbol{S}^{-1}\boldsymbol{x}_i$ と $MTA_i=\boldsymbol{x}_i'\boldsymbol{A}^*\boldsymbol{x}_i$ の比例関係は、前者が後者の

$1/|\boldsymbol{S}|=1.136364$ 倍という正比例の関係にある。

しかし、重要なことは(5.1)の随伴行列を求めなくても(5.1)の $\boldsymbol{S}$ から直接、逆行列が求められることである。変数間の相関が高い場合でも S がフルランクであれば $\boldsymbol{S}^{-1}$ が求められる。

本節の結論は、余因子行列を使っても解けない問題は解けない、解ける問題は余因子行列を使わなくても解ける、ということである。MTA 法に代わる方法はいくとおりかある。しかしこの問題は多重共線性への対応の問題なので 6.1 節であらためて整理する。

なお余因子行列の利用がなぜ発生したのかという時代的な背景を推察しよ

う。マハラノビス自身が 1936 年の論文でも分散共分散行列の余因子展開を示している。当時はコンピュータもなく、ごく小規模な問題を手計算していた時代であった。そのため、たとえば 4 次の相関行列を余因子展開して 3 次の行列式の計算に還元してサラスの方法を適用する手続きに意味があった。今日のコンピュータによる数値計算では、より大規模なデータに対して逆行列を求める高速なアルゴリズムが用いられている。その結果、余因子展開による逆行列の計算は今日的な意義を失っている。

5.2　コレスキー分解による汎距離の高速計算

後藤太郎はコレスキー分解によって汎距離の計算が速くなることを本節の実験で確かめた。$D^2=\left(\boldsymbol{x}-\boldsymbol{m}\right)'\boldsymbol{S}^{-1}\left(\boldsymbol{x}-\boldsymbol{m}\right)$という汎距離のケース 2 で、測定データをあらかじめ平均 0、標準偏差が 1 になるように規準化しておく。$\boldsymbol{m}=\boldsymbol{0}$ であるからサンプルi の群平均からの汎距離は(5.3)式で表される。なお$\tilde{\boldsymbol{x}}_i$ はサンプル i に関する p 次の列ベクトルである。変数ベクトルではなくてサンプルのベクトルであることをチルダで強調した。

$$D_i^2=\tilde{\boldsymbol{x}}_i'\boldsymbol{R}^{-1}\tilde{\boldsymbol{x}}_i \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.3)$$

測定データが規準化されていることから分散共分散行列 $\boldsymbol{S}$ は相関行列 $\boldsymbol{R}$ になる。$\boldsymbol{R}$ が実対称行列で正定値であるという条件のもとで、$\boldsymbol{R}$ を上三角行列$\boldsymbol{Q}$ の積に分解するのがコレスキー分解である。

$$\boldsymbol{R}=\boldsymbol{Q}'\boldsymbol{Q}$$

ここで $\boldsymbol{R}$ の逆行列をとれば　$\boldsymbol{R}^{-1}=\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}=\boldsymbol{Q}^{-1}\left(\boldsymbol{Q}^{-1}\right)'$　したがって(5.3)は

$$D_i^2=\tilde{\boldsymbol{x}}_i'\boldsymbol{Q}^{-1}\left(\boldsymbol{Q}^{-1}\right)'\tilde{\boldsymbol{x}}_i=\tilde{\boldsymbol{y}}_i'\tilde{\boldsymbol{y}}_i \quad \cdots\cdots\cdots\cdots\cdots \quad (5.4)$$

データ行列 $\boldsymbol{X}\left(n\times p\right)$ を使って n 人分をまとめて表記すれば

$$\underset{n\times p}{\boldsymbol{Y}}=\boldsymbol{X}\boldsymbol{Q}^{-1} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.5)$$

(5.5)の行列$\boldsymbol{Y}$ の第 i 行ベクトルを抜き出した列ベクトルが $\tilde{\boldsymbol{y}}_i$ である。(5.4)の右辺の最後の式はいかにも平方ユークリッド距離のように見えるが、実際は(5.3)の汎距離と同じ計算をしている。小さな数値でコレスキー分解を確かめよう。

$$\boldsymbol{Q}=\begin{bmatrix}1 & 0.8250503\\ 0 & 0.5650593\end{bmatrix},\ \boldsymbol{Q}^{-1}=\begin{bmatrix}1 & -1.460113\\ 0 & 1.769726\end{bmatrix},\boldsymbol{R}=\begin{bmatrix}1 & 0.8250503\\ 0.8250503 & 1\end{bmatrix}$$

$$\boldsymbol{Q}'\boldsymbol{Q}=\boldsymbol{R},\qquad \boldsymbol{Q}^{-1}\boldsymbol{Q}=\boldsymbol{I}$$

$\boldsymbol{Y}$ の列ベクトル $\boldsymbol{y}_1,\boldsymbol{y}_2$（こちらは変数ベクトル）を、上記の例で表現すれば

$$\boldsymbol{y}_1=\boldsymbol{x}_1,\qquad \boldsymbol{y}_2=-1.460113\boldsymbol{x}_1+1.769726\boldsymbol{x}_2$$

2 変数データを使って 1 万サンプルの汎距離計算を行った。同じテストを 1000 回繰り返して R のパッケージ ”microbenchmark” で計算速度を測った結果は次の通り。

①1 サンプルずつ定義通りに汎距離を計算した場合⇒平均 219.4 秒
②ベクトル演算で定義通りに汎距離を計算した場合⇒平均 0.6 秒
③(5.5)のコレスキー分解にもとづいて計算した場合⇒平均 0.5 秒

```
# コレスキー分解を使った汎距離計算用関数
MD_GramSchmidt <- function(data) {
  #汎距離計算用関数
  # グラムシュミット正規直交基底を使用
  X <- data
  X_bar <- colMeans(X) # 平均ベクトル
  Cov_X <- cov(X) # 分散共分散行列
  L <- chol(Cov_X) # コレスキー分解

  # グラムシュミット法による変換
  X_orth <- sweep(X, 2, X_bar) %*% solve(L) # コレスキー分解の逆行列を使って変換
  d2 <- rowSums(X_orth^2) # 汎距離の算出
  return(as.numeric(d2))
}
```

このテスト結果から、コレスキー分解を用いた汎距離計算はベクトル演算と同等の高速処理ができることが確かめられた。

5.3　閾値を計算するＲのコード

本節ではケース 2 の汎距離 D^2 が閾値を超えるかどうかを計算するコードを示す。棄却水準 α そのものは社会的な合意事項にすぎず、理論的に正しい α が与えられるわけではない。

MT 法ではカイ二乗分布を閾値計算のロジックに用いている。しかしそのためには変数が多変量正規分布に従うことが前提になる。現実のデータ解析では

D^2 は観測データから求めた統計量であってカイ二乗分布には従わない。閾値設定については宮川・永田（2022、pp124-132）が次の方法を示している。

観測データから計算した汎距離の推定値 $\hat{D}^2$ は次のベータ分布に従う。

$$\frac{1}{n-1}\hat{D}^2 \sim Beta\left(\frac{p}{2}, \frac{n-p-1}{2}\right) \qquad (5.6)$$

ベータ分布と F 分布の関係は第 1 自由度を ϕ_1、第 2 自由度を ϕ_2 として次の通りである。

$$Beta\left(\frac{\phi_1}{2}, \frac{\phi_2}{2}\right) = \frac{\frac{\phi_1}{\phi_2} F\left(\phi_1, \phi_2\right)}{1+\frac{\phi_1}{\phi_2} F\left(\phi_1, \phi_2\right)} \quad 、 \qquad (5.6)\text{より } \phi_1 = p, \quad \phi_2 = n-p-1 \text{ であ}$$

るから、F 分布の上側 100 α %点を閾値とした棄却域は次のように設定すればよい。

$$\hat{D}_2 \geq (n-1)\frac{\frac{p}{n-p-1} F\left(p, n-p-1; \alpha\right)}{1+\frac{p}{n-p-1} F\left(p, n-p-1; \alpha\right)} \qquad (5.7)$$

この閾値を計算する R のコードを以下に示す。ここで qf は F 分布の分位関数であり、累積分布の確率に対応した F 値を返す。alpha が上側確率を指すことを lower.tail = FALSE で指定している。

```
# 異常値判定のための閾値の設定の関数
MD_threshold <- function(n, p, alpha = .05) {
  #n: サンプル数
  #p: 変数の数
  #alpha: 棄却域（F分布,デフォルトは5%水準とする）
  f_stat <- qf(alpha, p, n-p-1, lower.tail = FALSE) #alpha 水準時の F 統計量
  thresh_d2 <- (n-1)^2/n * (p/(n-p-1) * f_stat/(1+p/(n-p-1) * f_stat)) # 閾値の計算
  return(thresh_d2) # 計算した閾値の表示
}
```

5.4　測定データの規準化

規準化の操作によって汎距離が変化するかどうかを検討しよう[7]。まず記法は次の通り。

$$V=\frac{1}{n}X'X, D^{\frac{1}{2}}=diag\left(s_j\right), Z=XD^{-\frac{1}{2}}, R=\frac{1}{n}Z'Z$$

平均偏差データ行列を X、変数 j の標準偏差を s_j と書いた。Sが分散を指すのか標準偏差を指すのかで混乱しないように本節では分散共分散行列を V で示した。Xは平均偏差化されているので、必要な規準化は $z_{ij}=x_{ij}/s_j$ である。この規準化を p 次のベクトルで書けば、$\tilde{\boldsymbol{z}}_i=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{x}}_i$ である。するとケース 2 のサンプル i と平均 0 との汎距離は

$$D_i^2=\tilde{x}_i'V^{-1}\tilde{x}_i=\tilde{x}_i'\left(\frac{1}{n}X'X\right)^{-1}\tilde{x}_i=n\tilde{x}_i'\left(D^{\frac{1}{2}}Z'ZD^{\frac{1}{2}}\right)^{-1}x_i$$

$$=nx_i'D^{-\frac{1}{2}}\left(Z'Z\right)^{-1}D^{-\frac{1}{2}}\tilde{x}_i=\tilde{x}_i'D^{-\frac{1}{2}}\left(\frac{1}{n}Z'Z\right)^{-1}D^{-\frac{1}{2}}\tilde{x}_i=\tilde{\boldsymbol{z}}_i'\boldsymbol{R}^{-1}\boldsymbol{z}_i$$

したがって汎距離はスケールフリーな尺度であることが分かる。このことは汎距離を原データのまま計算しても、規準化してから計算しても汎距離を用いた意思決定には影響しないことを意味する。

6 章 多変量解析と汎距離

6.1　回帰分析と多重共線性

回帰分析と汎距離には、変数間に一次従属の関係がある場合に分散共分散行列の逆行列が求められないという共通した問題がある。また回帰分析では安定的な解を得るために回帰診断や変数選択が研究されてきた。それらの手法は汎距離を利用する際にも利用できるだろう。

■重回帰分析のモデル

1 個の目的変数 $\boldsymbol{y}$ とp個の説明変数$\boldsymbol{X}$があり、n個の観測データが列に関して中心化（平均偏差化）されているとする。この時、重回帰分析のモデルは(6.1)で表せる。

$$\underset{n\times 1}{\boldsymbol{y}}=\underset{n\times p}{\boldsymbol{X}}\underset{p\times 1}{\boldsymbol{\beta}}+\underset{n\times 1}{\boldsymbol{\varepsilon}}, \quad \boldsymbol{\varepsilon}\sim N_n\left(\boldsymbol{0},\sigma^2\boldsymbol{I}_n\right) \qquad (6.1)$$

誤差(error) $\boldsymbol{\varepsilon}$ は平均がゼロベクトルで分散共分散行列が $\sigma^2\boldsymbol{I}_n$ のn変量正規分布に従う確率変数と仮定する。これは全ての $\varepsilon_j \quad (j=1,2,\cdots,p)$ が独立に同一の $N(0,\sigma^2)$ に従うと言っても同じである。(6.1)の確率変数は $\boldsymbol{\varepsilon}$ と $\boldsymbol{y}$ であり未知のパラメータは β である。最小二乗法を用いれば正規方程式 $\boldsymbol{X'X\beta = X'y}$ を解いて $\boldsymbol{\beta}$ の推定値 $\boldsymbol{b}$ が導かれる。

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \qquad (6.2)$$

この $\boldsymbol{b}$ を用いて $\boldsymbol{y}$ を予測する時の外れを残差(residual)ベクトル $\boldsymbol{e}$ と呼ぶ。$\boldsymbol{e}$ は観測値 $\boldsymbol{X}$ では説明できなかった情報を示す。

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}, \quad \boldsymbol{e} \sim N_n\left(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \Pi_X)\right) \qquad (6.3)$$

理論モデルである(6.1)の $\boldsymbol{\varepsilon}$ と統計量である(6.3)の $\boldsymbol{e}$ は異なる概念である。なお(6.3)の $\Pi_X = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ はp個の説明変数ベクトルで張られる部分空間への射影子である。$\boldsymbol{I} - \Pi_X$ は $\boldsymbol{y}$ を Π_X の直交補空間に射影する射影子である。

■罰則つき最小2乗推定量

$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$ のベクトルの組に一次従属の関係がある場合は(6.2)の逆行列が求められない。一次従属でなくても情報行列 $\boldsymbol{X'X}$ の行列式が 0 に近い場合は回帰係数の推定値が不安定になる。このようなたちが悪い(ill condition) データへの対策として、最小二乗基準に罰則項(正則化項ともいう)を加える提案がある。

よく知られた Ridge 回帰は罰則項を β の二乗和 $(\boldsymbol{\beta}, \boldsymbol{\beta})$ とおいて回帰係数を推定するものである。最小化すべき損失関数は未定乗数を ν として

$L(\boldsymbol{\beta}, \nu) = (\boldsymbol{e}, \boldsymbol{e}) + \nu(\boldsymbol{\beta}, \boldsymbol{\beta})$ でリッジ推定量は

$$\boldsymbol{b} = (\boldsymbol{X'X} + \nu I)^{-1}\boldsymbol{X'y} \qquad (6.4)$$

(6.4)を形式的に眺めれば、情報行列の主対角要素に小さな正数 ν を追加することで推定値を安定化させていると解釈できる。

また Tibshirani(1996)のLasso回帰は β の絶対値の和 $\sum_{j=1}^{p}|\beta_j|$ を罰則項にしてい

る。その損失関数 L は未定乗数を λ として次の通りである。

$$L(\boldsymbol{\beta},\lambda)=(\boldsymbol{e},\boldsymbol{e})+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

未定乗数が大きければ、損失関数を小さくするには β の絶対値の和を小さくしなければならない。その結果として 0 の値をとる回帰係数が増えることが期待される。Lasso回帰はデータ数 N が小さく変数の数pが大きいスパースデータの解析で活躍している。

■ 回帰診断

次の手順で回帰診断することが行われている。

1)行列式がゼロに近いかどうか

分散共分散行列 S、あるいは相関係数行列 Rの行列式を求めて、ゼロないしゼロに近ければ行列が ill condition であることが検出できる。問題の解決法は別にして問題があることだけは分かる。

2)CN(Condition Number)をチェックする

Sまたは Rの固有値を大きさ順に $\lambda_1,\lambda_2,\cdots,\lambda_p>0$と配列できたとして

$CN=\lambda_1/\lambda_p$ が大きければ問題ありと判断する。もしゼロの値の固有値が出た時は、一次従属の関係があると判定できる。

3)VIF(Variance Inflating Factor)が大きいか

j番目の説明変数を仮に目的変数に設定して、残りのp−1 個の説明変数を用いて重回帰分析を行って決定係数 r_j^2 を求める。$VIF_j=\dfrac{1}{1-r_j^2}$ である。j 番目の説明変数が他の変数と従属関係が高ければ r_j^2 は 1 に近くなるはずで、その結果 VIF は大きな値をとる。具体的な手順は VIF が 10 以上で最大の値をとる変数から順に一つずつ説明変数の組から除いて VIF を再計算するという方法がとられている。

■直交化予測法

説明変数を直交化させてから回帰分析を行うという対策もある。直交化にはコレスポンデンス分析、因子分析、数量化理論Ⅲ類など多くの方法が利用できるが、中でも主成分分析がよく使われる。主成分分析は分析上のハイパーパラメータが少ないので、誰が分析しても同じデータから同じ主成分スコアを再現しやすいからである。もちろんプログラムが間違っているため再現できないこともあるので安心はできないが。

空間を直交化させてもすべての次元を説明変数に用いて回帰分析をするのであれば、決定係数は改善されない。なぜならp変数もp主成分も同じ空間の基底なので同じ決定係数をもたらすからである。

そこで直交化の途中の$r(<p)$で次元を打ち切ることにすれば、説明変数空間が違ってくる。p変数よりも次元を縮小した基底を予測に用いることは、実務の上でも価値がある。

6.2　多重共線性への対策

松本健は本節で次の2つの問題を検討した。

1）正確多重共線性がある場合、汎距離をどう求めればよいか
2）準多重共線性の場合にはどういう対処ができるか

多重共線性(multicollinearity)というのはあいまいに使われている用語なので、まず図 6.1 のように概念を整理しよう。

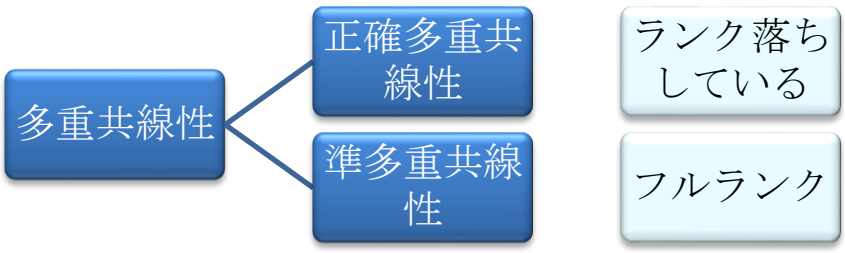


図 6.1　多重共線性の整理

一次独立な変数の組の最大数をランクといい、変数の数よりランクが少ない場合をランク落ちという。たとえば p 次の分散共分散行列 S の$rank(\boldsymbol{S}) < p$　であれば、その変数の組は正確多重共線性がある[8]。一方準多重共線性とは$rank(\boldsymbol{S}) = p$　ではあるもののSの行列式が 0 に近い状態をいう。この近いというのは、どれだけ近ければ「近い」のかという基準はあいまいである。

■　一般逆行列を使う

正確多重共線性がある$\boldsymbol{S}$については逆行列が求められない。その対策はいくつか考えられるが、Srivastava(2006)の提案に従って一般逆行列を適用してみよう。

柳井・竹内(1983）によれば一般逆行列は４種類あるが、本節ではムーアペンローズの一般逆行列を用いる。その理由は、回帰分析の文脈で言えば、誤差の最小二乗、回帰係数の最小ノルムかつ一意な解が得られるからである。

データの作成法を R のコードで示す。変数 X1~X5 が独立な標準正規分布に従うとして、各変数ごとに 300 個の乱数を発生させてデータ行列 df.1 を作る。

さらに X1~X5 の合計 Xsum を追加することで一次従属なデータ行列 df.2 を作成した。

```
N <- 300
x1 <- rnorm(N)
x2 <- rnorm(N)
x3 <- rnorm(N)
x4 <- rnorm(N)
x5 <- rnorm(N)
x_sum <- x1+x2+x3+x4+x5
df.1 <- cbind(x1, x2, x3, x4, x5)
df.2 <- cbind(x1, x2, x3, x4, x5, x_sum)
```

今回の実験では 6 つの変数の相関行列は表 6.1 の通りであった。非主対角要素に相関係数 1.0 は存在しないにもかかわらず $rank(df.2) = 5$ と df.2 のランクは落ちている。このことは、多重共線性を発見するには相関行列を点検すれば大丈夫だという実務ノウハウが誤りであることを示している。

表 6.1　6 変数の相関行列

	x1	x2	x3	x4	x5	x_sum
x1	1.000	-0.061	-0.045	0.016	-0.120	0.332
x2	-0.061	1.000	-0.040	0.112	-0.014	0.445
x3	-0.045	-0.040	1.000	0.040	0.028	0.458
x4	0.016	0.112	0.040	1.000	0.079	0.568
x5	-0.120	-0.014	0.028	0.079	1.000	0.430
x_sum	0.332	0.445	0.458	0.568	0.430	1.000

逆行列と一般逆行列を用いてケース 2 の汎距離を計算したところ表 6.2 の結果が得られた。

表 6.2　汎距離の計算出力

利用する関数	フルランクの $df.1$	ランク落ちした $df.2$
逆行列の関数　　solve( )	出力 1	計算不能
一般逆行列の関数 ginv( )	出力2	出力3

一般逆行列を使えばフルランクだけでなくランク落ちした $\boldsymbol{S}$ に対しても逆行列が求められる。さらに重要なことはたんに数値が出るだけでなく、どちらからも同一の距離が得られることである。以下にデータの一部を示すが、表 6.2 の 3 つの出力はすべて数値が一致する。このことは一般逆行列で汎距離計算を行うプログラムを用意すれば、変数の組をフルランクに縮減した場合と同じ汎距離が求められることを意味している。その意味でも一般逆行列は通常の逆行列を一般化している。

	出力 1	出力 2	出力 3
[1,]	3.232637	3.232637	3.232637
[2,]	1.830617	1.830617	1.830617
[3,]	4.501335	4.501335	4.501335
[4,]	7.534101	7.534101	7.534101
[5,]	2.167525	2.167525	2.167525

（以下、省略）

■ 回帰分析の変数選択

準多重共線性のあるデータを作ってテストしてみよう。回帰分析のモデルをとりあげて、X1 と X2 が類似した条件で偏回帰係数がどう推定されるかを調べる。この実験では X1 に正規乱数を加えて X2 を作成した。X1 と X2 の相関係数は 0.919 である。

```
x1 <- rnorm(N)
x2 <- x1 + 0.4*rnorm(N)      #   x1 と x2 は相関が高い
x3 <- rnorm(N)
e <- rnorm(N)
y   <- 1000 + 100*x1 + 80*x2 + 60*x3 + 100*e
df.3 <- data.frame(x1, x2, x3, e, y)
# 重回帰分析
model_lm3 <- lm(y~x1+x2+x3, data=df.3)
summary(model_lm3)
```

分析結果は次の通りで、一次従属に近い変数 X1 と X2 の偏回帰係数は標準誤差が大きく、それぞれのt値は小さくなる。

【3 変数の分散分析表】

	係数	標準誤差	t 値	p値
定数	1002.441	5.874	170.650	< 2e-16 ***
x1	73.000	15.791	4.623	5.66e-06 ***
x2	109.728	14.904	7.363	1.80e-12 ***

x3	64.558	5.700	11.327	< 2e-16 ***

6.1 節で紹介した VIF を使って変数選択を行うと、X1 と X2 を一緒に説明変数に入れた予測モデルの VIF は高く、それ以外の組み合わせでは VIF は小さくなる。この回帰診断を利用して X1 と X3 を選択した結果は次の通りである。

【2変数の分散分析表】

	係数	標準誤差	t 値	p値
定数	1002.970	6.378	157.25	<2e-16 ***
x1	179.855	6.757	26.62	<2e-16 ***
x3	62.766	6.183	10.15	<2e-16 ***

X1 の偏回帰係数の標準誤差は 3 変数のときより小さくなり、逆にt値は大きくなった。準多重共線性を回避することで推定結果がより安定的になった。回帰分析では他にも変数の組み合わせによる F 値の変動をみるなど、変数選択の手段が研究されている。

実験の結論は次の通りである。多重共線性の可能性がある事態では、次の対策が有効であった。

1)変数をそのまま用いて一般逆行列で汎距離を計算する

2)VIF を利用して変数を減らしてから通常の逆行列で汎距離を計算する

6.3　群のサイズを考慮した判別分析

■ 線形判別関数

線形判別関数は次のフィッシャー(1936)の仮定に従うことで尤度比基準から導くことができる。なお本節での記法は(3.2)、(3.3)に従う。

【Fisher の仮定】集団 1,2 がそれぞれ多変量正規分布に従うと仮定する。さらに説明変数 $\mathbf{X}$ の平均ベクトル $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ は 2 群で異なるが、分散共分散行列 $\boldsymbol{\Sigma}$ は両群で等しいと仮定する。

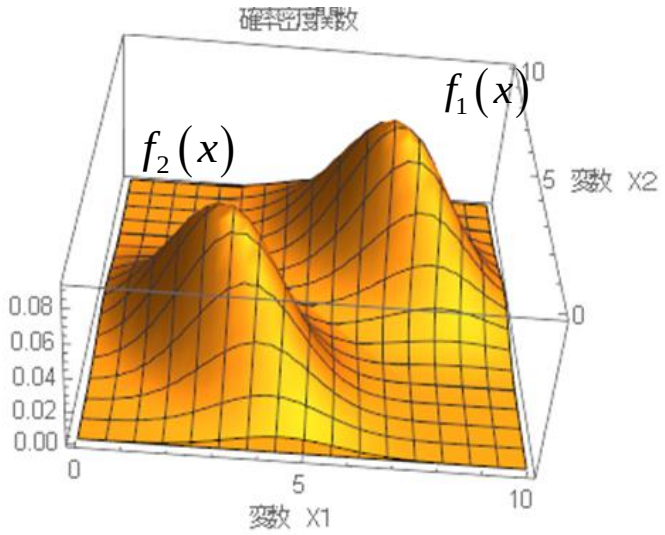


図 6.2 2群の多変量正規分布

2 群の確率分布を描けば図 6.2 のようになる。図 6.2 は $p=2$ の場合を図示したものである。$\mathbf{x}$ を何らかの観測データに固定すれば、$f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$ はパラメータ $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ の尤度（ゆうど）を表す。尤度比から次の(6.5)式が導かれる。

$$
\begin{aligned}
\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] \\
&= \exp\left[\frac{1}{2}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\}\right] \\
&= \exp\left[\frac{1}{2}\left\{\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}-\boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_2-\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{x}+\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{\mu}_2-\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}+\boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{x}-\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1\right\}\right] \\
&= \exp\left[\frac{1}{2}\left\{-2\boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_2+2\boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{\mu}_2-\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1\right\}\right] \\
&= \exp\left[\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right] \\
&= \exp\left[\mathbf{x}'\mathbf{b}-\mathbf{a}'\mathbf{b}\right] \qquad (6.5)
\end{aligned}
$$

(6.5)の最後では次のようにベクトルを要約した。

$$\mathbf{a}=\tfrac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)、\quad \mathbf{b}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

ここで(6.5)右辺の指数部に注目すると

$$h(\mathbf{x})=\mathbf{x}'\mathbf{b}-\mathbf{a}'\mathbf{b} \qquad (6.6)$$

これがフィッシャーの線形判別関数 linear discriminant function であった。**b** は判別係数と呼ばれる。

(6.5)のパラメータを観測値に置き換え、$\boldsymbol{x}$ に測定値を代入した時の(6.6)の値が判別スコアである。$\exp 0 = 1$ であるから(6.6)が正の時に(6.5)は 1 より大きくなる。したがって、$\boldsymbol{x'b} > \boldsymbol{ab}$ であればこのサンプルは群 1 に属し、それ以外は群2に属すると判定すればよい。(6.5)の尤度比の分子と分母を入れ替えても、線形判別関数の符号が反転するだけで判別の結論は変わらない。

線形判別関数はマーケティングの実務でもよく利用されているが、判別の境界が直線であることからくる限界がある。2 群が超球状ではなく非球状に分布する場合に問題が生じる。しかも 2 群の分散共分散行列が異なって、シンプソンズパラドックスが該当する場合は、線形判別関数のモデルと現実との乖離が大きくなる。図 6.3 にその典型的な場合を示した。図の直線のように空間を分けても誤判別が多く発生することは避けられない。

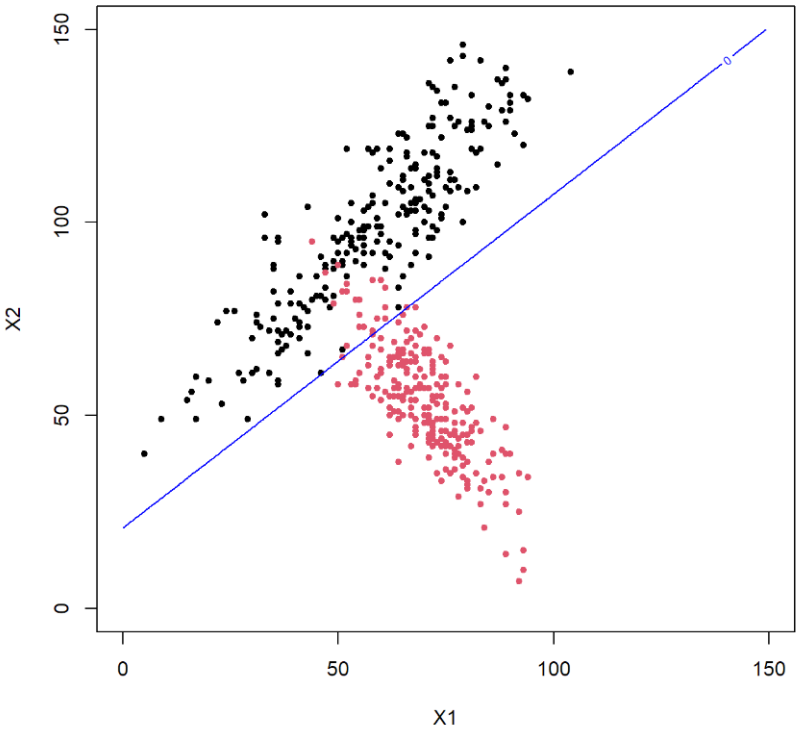


図 6.3　線形判別関数による境界線では上手く判別できない例

■ 群のサイズの違いを取り入れる

線形判別関数に群サイズの情報を取り入れることは次のように行われてきた。[9]群1の出現確率を θ とすれば、群の分割が排反で悉皆とすれば群2の出現確率は $1-\theta$ になる。

群 1 である可能性が群2である可能性を上回るのは $f_1(x)\theta > f_2(1-\theta)$ の場合だと考えてよいから

$\dfrac{f_1(x)\theta}{f_2(x)(1-\theta)} > 1$ が成り立つ場合である。左辺の対数をとれば

$$Q = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\theta}{1-\theta} \qquad (6.7)$$

一方右辺の対数は $\log 1 = 0$ なので、$Q > 0$ のときに $\boldsymbol{x}$ のサンプルは群 1 に属すると判定すればよい。

■フィッシャーの仮定をゆるめる

分散共分散行列が 2 群で異なる場合は、各群の S_1, S_2 を区別して扱わなければならない。この場合は、ケース6の汎距離 D^2 を用いて判別すればよい。

まず $\dfrac{f_1(x)}{f_2(x)}$ を測定データから計算すると(6.8)の通りになる。

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{|S_1|^{-1/2}}{|S_2|^{-1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-m_1)' S_1^{-1}(\boldsymbol{x}-m_1) + \frac{1}{2}(\boldsymbol{x}-m_2)' S_2^{-1}(\boldsymbol{x}-m_2)\right]$$

$$= \frac{|S_2|^{1/2}}{|S_1|^{1/2}}\exp\left[\frac{1}{2}\left\{D_2^2(x,m_2) - D_1^2(x,m_1)\right\}\right] \qquad (6.8)$$

$$= \sqrt{\frac{|S_2|}{|S_1|}}\exp\left[\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2|g_2) - D^2(\boldsymbol{x},\boldsymbol{m}_1|g_1)\right\}\right]$$

次に(6.7)に従って

$$Q = \log\sqrt{\frac{|S_2|}{|S_1|}} + \frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2|g_2) - D^2(\boldsymbol{x},\boldsymbol{m}_1|g_1)\right\} + \log\frac{\theta}{1-\theta} \qquad (6.9)$$

(6.9)の Q が正のときにそのサンプルを群1に判定する。(6.9)の判別法は汎距離によって解釈したものである。実際の計算内容は従来、2 次判別関数と呼ばれてきたものと一致する(柳井他、1986, 132 頁)。なお $S_1 = S_2$ の場合は(6.9)の第1項はゼロとなって消え、2群のサイズの違いを示す第 3 項が Q に残ることになる。

(6.9)を R の関数でコードしたのが次の囲みである。

```
two_class_qda <- function(data, class) {
    # data: 説明変数行列（class は含まない）
    # class: 分類クラスベクトル（2 群）

    # 群別統計量
    two_class_list <- split(data, class)
    two_class_mu <- lapply(two_class_list, function(x) colMeans(x))
    two_class_Sinv <- lapply(two_class_list, function(x) solve(cov(x)))
    two_class_det <- lapply(two_class_list, function(x) det(cov(x)))
    two_class_n <- lapply(two_class_list, nrow)
    theta <- two_class_n[[1]]/nrow(data)
    # 距離計算
    X <- data.matrix(data)
    d_1 <- rowSums(sweep(X, 2, two_class_mu[[1]]) %*% two_class_Sinv[[1]] *
sweep(X, 2, two_class_mu[[1]]))
    d_2 <- rowSums(sweep(X, 2, two_class_mu[[2]]) %*% two_class_Sinv[[2]] *
sweep(X, 2, two_class_mu[[2]]))
    # 判別関数の計算
    score <- log(sqrt(two_class_det[[2]]/two_class_det[[1]])) + 1/2*(d_2-d_1) +
log(theta/(1-theta))
    # 判別結果の分類
    class <- ifelse(score > 0, "group_1", "group_2")
    # 結果の格納
    return(list(score = score,
                class = class))
}
```

■ 群に帰属する予測確率を求める

(6.9)から 2 群のどちらに判定するかという 2 値の出力が得られた。それに対して、本項では確率的な予測を行う方法を述べる。

θ はサンプルから得られる情報とかかわりがない情報なので事前確率と呼ばれる。次にベイズの定理を使ってデータが $\boldsymbol{x}$ であるサンプルがg1 に所属する確率 P を求める。

$$
\begin{aligned}
&P\left(g_1|\mathbf{x}\right) \\
&= \frac{f_1(\mathbf{x})\theta}{f_1(\mathbf{x})\theta + f_2(\mathbf{x})(1-\theta)} = \frac{1}{1+\dfrac{1-\theta}{\theta}\cdot\dfrac{f_2(\mathbf{x})}{f_1(\mathbf{x})}} \\
&= \frac{1}{1+\exp\left[\log\dfrac{1-\theta}{\theta}+\log\dfrac{f_2(\mathbf{x})}{f_1(\mathbf{x})}\right]}
\end{aligned}
\tag{6.10}
$$

予測確率は(6.8)、(6.9)の結果を用いて次のように求められる[10]。

$$
\begin{aligned}
&P\left(g_1|\mathbf{x}\right) \\
&= \frac{1}{1+\exp\left[\log\left(\dfrac{1-\theta}{\theta}\sqrt{\dfrac{|S_1|}{|S_2|}}\right)+\dfrac{1}{2}\left\{D^2\left(\boldsymbol{x},\boldsymbol{m}_1|g_1\right)-D^2\left(\boldsymbol{x},\boldsymbol{m}_2|g_2\right)\right\}\right]} \\
&= \frac{1}{1+\exp(-Q)}
\end{aligned}
\tag{6.11}
$$

近年では一般の消費者だけではなく、発生が極めて希なマーケット、たとえば超富裕層のマーケティングも必要になっている。希少ターゲットの場合は、5 分 5 分の基準で見込み客かどうかを判別するのは適切ではないだろう。場合によっては $P\left(g_1|\mathbf{x}\right)>0.01$ でも個別対応する必要があるかもしれない。その意味でマーケティングの実務では、確率的な予測ができる(6.11)の方が 1 か 0 かの判定よりも利用価値があるだろう。

6.4　クラスター分析と汎距離

■　クラスター分析と距離についての問題

マーケティングリサーチの実務では、消費者をセグメントする際に、あらかじめ原データを因子得点や主成分得点に変換してからクラスター分析することが行われてきた。因子分析の主因子解および主成分は直交しているので、ユークリッド距離で消費者間の距離を測ってよいはずだ、というのが従来の通説だったようである。

しかし空間が複数の集団から成り、しかも集団によって分散共分散行列が異なるという一般的な事態に、はたしてこの理屈は通用するのだろうか。なぜなら、データを一括して直交化したとしても、個々の集団内の分布まで無相関になるわけではない。したがって直交化後のユークリッド距離と集団ごとに測っ

た汎距離は一致しない。集団が複数の場合の距離の測定はこれまで見過ごされてきたのではないだろうか。本節ではクラスターごとに汎距離を測る方法を検討する。

■　ノンパラメトリックな立場

クラスター分析には正規混合法(GMM)のようなモデルベースのクラスター分析もあるが、現実のデータにモデルベースの理論が適合する保証はないので本節ではノンパラメトリックなクラスター分析を扱う。

クラスター分析は大きく分ければ、距離の小さいサンプルどうしを段階的に併合して大きなクラスターにまとめる階層クラスター分析と、N 個のサンプルを K 個のクラスターに分割する非階層クラスター分析に分けられる。

どちらのクラスター分析でも、これまで汎距離はあまり利用されてこなかった。Cerioli(2005) はその理由として、全ての点が 1 つの集団 single population から来ているという汎距離の定義が、空間は複数の集団からなるというクラスター分析の認識と相反するからだとしている。[11] とくに階層クラスター分析においては、空間の任意の 2 点間の汎距離を測る必要があるが、4.5 節で指摘したように、任意の 2 点間の汎距離の計測は容易ではない。

さらに実務的な観点をいえば、マーケティングでは少数のデータよりも大規模なデータをクラスター分析したいというニーズが高いことがあげられる。ビッグデータのデータマイニングがその例である。

そこで本節では、大規模データに適した非階層型のクラスター分析をとりあげる。そしてクラスター分析に汎距離を導入することでユークリッド距離よりもパフォーマンスが高まるかについての既存研究を紹介する。

■　非階層クラスター分析

MacQueen (1967) の k-means 法は非階層クラスター分析の代表的な方法である。しかしベーシックな k-means 法は初期値の選択に欠点があった。乱数でサンプルを選んでシード（クラスターの核）にする方法なので、抽出次第では多次元空間の一部の領域にシードが集中する危険があった。またシードに依存して反復解が局所的最適値に収束しがちなため分析結果の再現性が低く信頼性に欠ける方法だった。

セリオリによれば、k-means 法はデータが超球状に散布しており、各クラスターのサイズもほぼ等しいことを暗黙に仮定した方法であるという。ということは、真のクラスターが非球状 non-spherical の形をしているときに k-means 法を使うのは不適切だということになる。k-means 法に対してはいくつも改訂案が出されているので、主要な提案を紹介しよう。

■　汎距離の導入

アンドレア・セリオリが 2005 年に提案した Modified convergent k-means

algorithm を次に示す。

1）ユークリッド距離を用いてK（$k=1,2,\cdots,K$）個のシードを選んで、それらを暫定的にセントロイドにする。セントロイドとは集団の重心や中心の意味である。シードは careful に選ぶ。

2）$i=1,2,\cdots,n$ 個のデータを、ユークリッド距離が一番小さい暫定セントロイドのグループに入れて初期クラスターを作る。ここでは、N 個のサンプルすべてをクラスターに所属させる必要はない。初期クラスターのサイズ cs は適当に定めればよい。

3）各クラスターの平均ベクトル $\boldsymbol{m}_k$ と第 k クラスターの分散共分散行列 $\boldsymbol{S}_k$ を計算する。クラスターk のサンプル数を n_k と書く。ただし n_k は少なくとも変数以上でないと $\boldsymbol{S}_k$ は計算できない。

4）サンプル i $\left(i=1,2,\cdots,N\right)$ と平均ベクトルの間の、我々がいうケース 6 の汎距離を計測する。 $D^2\left(\boldsymbol{x}_i,\boldsymbol{m}_k\right)=\left(\boldsymbol{x}_i-\boldsymbol{m}_k\right)' S_k^{-1}\left(\boldsymbol{x}_i-\boldsymbol{m}_k\right)$

セリオリは上式の平方根をとった D を用いている。しかしクラスターへの所属判定は D, D^2 のどちらでも同じである。なお逆行列が求められない時はユークリッド距離 d, d^2 を使う。D ならｄを使うというように汎距離とユークリッド距離のベキ乗を合わせればよい。

5）サンプル i の所属変更

サンプル i をDの値が一番小さいクラスターに所属替えする。n 個のサンプルについて変更が済んだら $\boldsymbol{m}_k$ と $\boldsymbol{S}_k$ を再計算する。

6）上の4）と5）の計算をイテレーションする。イテレーションの番号をｔとする。

7）所属クラスターの変更が無くなれば $\boldsymbol{m}_k$ と $\boldsymbol{S}_k$ はイテレーションの前後で変化しなくなるので収束と判断する。あるいはｔがユーザーが指定した上限 Tstop に達したらイテレーションを打ち切る。

セリオリのアルゴリズムでユーザーが指定するハイパーパラメーターは K, ns, Tstop の 3 つである。

セリオリのシミュレーション実験によれば、セリオリの方法による誤分類の

率 R はゼロに近く、ベーシックな k-means 法や SPSS の k-means 法よりも優れていた。フィッシャーのアヤメのデータで 4 変数、K=3 で実行したところ、$R = 5/150 = 0.033$ という結果になった。結論としてユークリッド距離よりも汎距離の優秀さが実証された。セリオリは他にも次のような指摘をしている。

■各クラスターの $\boldsymbol{S}_k$ が等しいという pooled $\boldsymbol{S}_k$ という仮定は非合理的である。4 章の(4.1)でこのプーリング法の計算式を示した。

■データセットのサンプルをどう並び替えても、番号名は別にして実質的に同じクラスター分割が導かれることが望ましい。

データ入力の order effect がないこと及び収束が早いことはどちらもクラスター分析の評価指標である。もし教師信号にあたる真のグループが既知の場合は、その情報を隠してクラスター分析を行い真のグループを予測できたかを評価することができる。

さてセリオリのステップ１）で、どうシードを careful に選ぶのかが残る課題になる。そこで Arthur ら(2007)は次のシード選択を提案した。

■ Careful Seeding

k-means 法の初期値を選ぶにあたって、単純に一番遠く離れたサンプルを選んでしまうと、外れ値もシードに選ばれてしまい、その近くにはデータポイントが何もないという事態が起きる。もちろん既存のシードから離れるのはよいことだが、外れ値は避けたい、それがアーサーらの k-means 法++であった。

アーサーの上手い工夫は各サンプルと最近隣シードとの距離 d(x)を測ったことにある。シードが１つしかない段階では、そのシードとの距離が d(x)である。シードが 2 つになったら、両方のシードとサンプルの距離を測って小さい方の値を d (x)に使う。以下シードが増えても同様である。シードもセントロイドの一種だと考え、以下ではセントロイドという用語で記述する。データ行列を $\boldsymbol{X}$ とする。

【k＋＋のアルゴリズム】

1a. 1 番目のセントロイドを $\boldsymbol{X}$ から一様乱数で選ぶ

1b. 次のセントロイドは次のウェイトを用いて確率的に $\boldsymbol{X}$ から選ぶ

$$w_i = \frac{d\left(x_i\right)^2}{\sum_{i=1}^{n} d\left(x_i\right)^2} \qquad (6.12)$$

1c. ステップ 1b をK個のセントロイドが選ばれるまで繰り返す

2－. そこから先はセリオリのアルゴリズム２）以降と同じである

なお (6.12) のウェイト付けをアーサーは「D2 ウェイティング」と呼んだ。しかし、D2 では汎距離と混同されるおそれがある。d(x)は最短の

セントロイドとのユークリッド距離なので、むしろ$d\min(x_i, c_k)$ のような記法にした方が説明的だったろう。

アーサーらが(6.12)で距離を平方した意図は、距離の違いを強調させる狙いがあった。その結果、既存のシードのそばのデータポイントが次のシードに選ばれる確率はゼロに近くなる。d(x)が小さければ選択確率が下がるからである。

本当に望ましい k-means 法の初期値は、初めからデータが集積している塊を発見してそこからシードを選ぶことである。夜空の銀河系を眺めればいくつもの星雲が見える。そこから一つずつ代表の星を選べば済む話である。低次元の空間なら簡単なことでも、空間の次元が数千あるいは数万になると人間が集積場所を目で見て判断するのは難しい。

そこでネルソン (2012) はデータが近傍に多くあるデータポイントをシードに選ぶことを提案した。これはもっともなアイデアなのだが、ネルソンのアルゴリズムは組み合わせ計算のコストが大きい。そのため実務で使われたという事例はまだなさそうである。

7章　討論

7.1　本研究で明らかになったこと

(1) マハラノビスの汎距離は同名異義で使われてきた

マハラノビスの汎距離は研究者によって定義が異なることがあった。たとえば MT システムでいうところのマハラノビスの汎距離はマハラノビスが提唱した汎距離とは異なる。本研究では汎距離を 7 種類に区別し、それらの関係を整理した。

適切な用語法としてはマハラノビスが提唱した汎距離はマハラノビスの汎距離と呼び、原案に触発されてその後派生した汎距離は単に汎距離と呼ぶのがよいのではないだろうか。汎距離は言葉で説明するよりも数式で計算法を示す方が簡明である。

(2) 集団が１つの場合、平均からの汎距離はスケールフリーである

集団が 1 つだけの場合は、測定変数の単位を変更しても汎距離は invariant(不変)である。それに対してユークリッド距離は測定単位を変えれば数値が変る。尺度としては汎距離の方が優れている。

しかし市場が複数の集団から構成される場合は、集団ごとに平均だけでなく分散共分散も変わるだろう。したがって集団ごとに違った汎距離を定義する必要がある。つまり汎距離は集団に条件づけられて決まる尺度である。原データを規準化したとしても問題は解決されない。

(3) 多重共線性への対処には一般逆行列が利用できる

MT システムの一技法である MTA 法では、測定変数の中に多重共線性の関係があった場合でも分散共分散行列（または相関行列）を余因子展開することで問題が解決できるとしていた。しかしランクが落ちた行列は余因子展開しても逆行列は導けない。余因子への評価をまとめると次の通り。

逆行列が解ける問題なら余因子を使わなくても解ける 逆行列が解けない問題は余因子を使っても解けない

本研究では正則なデータに一次従属な変数を追加して特異行列にしたケースについて次の性質を確認した。

■正則でないデータでもムーアペンローズの一般逆行列で汎距離が計算できる。

■しかもその汎距離は元の正則なデータから求めた汎距離と一致する。

（4）希少マーケットの発見

分散共分散行列が 2 群間で異なっていても、ケース 6 の汎距離で判別分析はできる。さらに 2 群への判定だけでなく、関心のある群への帰属確率を個人別に推定できる。

近年では超裕福層のような希少マーケットをターゲットにしたマーケティングが課題になることがある。そのような市場については、2 群のどちらかという判別分析は有効ではない。なぜならレア群の該当者はほぼいないという結論になりがちだからである。むしろレア群である見込み確率が 0.1 である、というような情報をもとに個別対応するのが実践的だろう。

(5) 2 群の平均値間の距離の一般化

マハラノビス自身は D^2 統計量の定義にあたって分散共分散行列が 2 群間で等しいという制約をおいた。この制約を外した平均値間の距離については、先行研究はまだないようである。我々は本研究で次の統計量を提案した。群番号を $h \neq g$ として

$$D^2\left(\boldsymbol{m}_h, \boldsymbol{m}_g \middle| G=g\right)=\left(\boldsymbol{m}_h-\boldsymbol{m}_g\right)' S_g^{-1}\left(\boldsymbol{m}_h-\boldsymbol{m}_g\right)$$

この距離は $d(x,y)=d(y,x)$ という対称性を満たさないので、公理論的な立場からすれば疑距離というべきである。しかしマーケティングにおいては、プレーヤー間に非対称な関係が存在することは珍しくない。たとえばサプライチェーンを構成する企業の関係がそうであるし、企業と顧客の間、そして顧客集団の間にも非対称性が起こり得る。平均間の距離になぜ非対称性が生じるかについてはグラフィカルな説明が可能である。

7.2　今後の研究課題

6章では回帰分析と判別分析に汎距離が深くかかわっていることを確認した。しかしクラスター分析については研究の途上にある。機械学習の言葉でいえばクラスター分析には教師データがない。クラスター分析の結果を評価する基準もはっきりしていない。

今のところクラスター分析では k-means++がよく利用されている。提唱者であるアーサーの D2 ウェイティングは距離と単調増加の関数をウェイトにしてクラスターのシードを逐次抽出する方法であった。D2 ウェイティングによって既存のシードに近い点を次のシードに選ぶ可能性は抑制できた。しかしシードからの最遠点ほど他のデータポイントより選択されやすいという仕組みなので、まだ改良の必要性はある。たとえばディープラーニングの活性化関数と同様に、シード抽出にシグモイド関数を利用するという修正方向も考えられる。今後の研究課題としたい。

クラスター分析には市場がいくつのクラスターから成っているのかが不明であるという根本的な問題がある。統計学的な基準からクラスター数を決める方法については豊田ら(2011)の研究がある。

汎距離はマーケティングのさまざまな評価システムにも、その構成要素(ingredient)として貢献できる可能性があるだろう。ブランド診断や優良顧客の識別などマーケティングに出てくる変数には変数間に相関があることが珍しくない。今後、現実のマーケティング課題に汎距離を適用し、その効果を実証していく必要がある。

謝　辞

本ワーキングペーパーのプログラムのコードを書いてくれた松波成行、松本健、後藤太郎の 3 氏に感謝する。

引用文献

Arthur, D. and Vassilvitskii,S. (2007) k-means++: the advantages of careful seeding. SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms , 1027-1035.

Cerioli,A (2005) K-means cluster analysis and Mahalanobis metrics：A problematic match or an overlooked opportunity?. *Statistica Applicata*, **17**,(1),61-73.

Fisher,R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**,179-188.

Harville,D.A.(1997) *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York. 邦訳：伊理正夫監訳（2007）「統計のための行列代数　上」シュプリンガー・ジャパン

Koschnick,W.J.(1996) *Dictionary of Social and Market Research.* John Wiley & Sons, Inc.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability, 281-297.

Mahalanobis,P.C.(1930) On test and measures of group divergence: theoretical formulae. *Journal of the Asiatic Society of Bengal*,**26**, No.4,541-588.

Mahalanobis,P.C. (1936) On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India, **2**(1),49-55.

Mahalanobis,P.C. (1949) Historical note on the D^2-statistic. *Sankhya*,**9**,237.

宮川雅巳・永田靖 (2022)「タグチメソッドの探求ー技術者の疑問に答える 100 問 100 答」日科技連

水野欽司（1996）「多変量データ解析講義」朝倉書店

永田靖・棟近雅彦（2001）「多変量解析法入門」サイエンス社

永田靖（2005）「統計学のための数学入門 30 講」朝倉書店

Nelson,J.D. (2012) On K-means clustering using Mahalanobis distance. Master thesis, North Dakota State University.

奥野忠一・他（1971）「多変量解析法」日科技連

Srivastava,M.S. (2006) Minimum distance classification rules for high dimensional data. *J. Multivariate Analysis,* **97**,2057-2070.

Taguchi,G., Chowdhury,S. and Wu ,Y.(2000) *The Mahalanobis-Taguchi System.* McGraw Hill Professional.

Taguchi,G., and Jugulum,R. (2002) *The Mahalanobis-Taguchi Strategy:A pattern technology system.* John Wiley & Sons, Inc.

田口玄一（1999）「品質工学の数理」日本規格協会

田口玄一・兼高達貳 編(2002) 「MT システムにおける技術開発」日本規格協会〈品質工学応用講座〉

田口玄一（2002）20 世紀の MTS 法と 21 世紀の MT 法、標準化と品質管理、55, [2],61-70.「品質工学の数理」日本規格協会
竹内啓・柳井晴夫（1972）「多変量解析の基礎」東洋経済新報社
田中豊・脇本和昌（1983）「多変量統計解析法」現代数学社
立林和夫編著（2008）「入門 MT システム」日科技連
Tibshirani,R.（1996）Regression shrinkage and selection via the lasso. *J. Royal Statistical Society, Ser B* ,**58** (1),267-288.
豊田秀樹・池原一哉(2011) 変数間の関係性を考慮してクラスター数を決定する k-means 法の改良、心理学研究,**82** ,(1),32-40.
Truett,J.,Cornfield,J. Kannel,W. (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Disease*, **20**, 511-524.
Venables,W.N.,Ripley, B.D. (2002) *Modern Applied Statistics with S* .4th ed., Springer.
Wedel,M.,Kamakura,W.(2000) *Market Segmentation.* Kluwer Academic Publishers.
柳井晴夫・竹内啓（1983)「射影行列・一般逆行列・特異値分解」東大出版会
柳井晴夫（1994）「多変量データ解析法－理論と応用－」朝倉書店
柳井晴夫・高木廣文編著（1986）「多変量解析ハンドブック」現代数学社
Yonenaga,K. and Suzukawa,A. (2021) Bayesian estimation for misclassification rate in linear discriminant analysis. *Japanese Journal of Statistics and Data Science*, **4**, 861-885.

【研究会の実施概容】
研究会のメンバーは次の通りである。

研究代表者　朝野熙彦　東京都立大学元教授
研究メンバー
　　奥瀬喜之　専修大学
　　藤居　誠　城西国際大学
　　河原達也　東京経済大学
　　松波成行　物質・材料研究機構
　　後藤太郎　CCC
　　松本　健　メルカリ
　　Deddy Jobson メルカリ
　　田村　覚　三井物産
　　朱　昱　　クロスマーケティング
サポーター
　　梅山貴彦　JMRA リサーチイノベーション委員会委員長

研究期間：2022 年 6 月 15 日～2023 年 3 月 8 日（月 1 回）
研究会場：専修大学神田キャンパス

---

1 マハラノビスはインド統計研究所 Indian Statistical Institute を 1931 年に設立した。彼がフィッシャーを 1937 年に同研究所に招いたことが、竹内啓(2018)「歴史と統計学ー人・時代・思想」日本経済新聞社の 264 頁に記述されている。Taguchi & Jugulum (2002)によれば田口も 1954 年に同研究所に研究員として招かれている。

2 田口玄一(1999)「タグチメソッドわが発想法」経済界で紹介されている。ここでの文章は木村浩訳（1972）トルストイ「アンナ・カレーニナ」新潮文庫の第 1 編から引用した。

3 集合には内包的定義と外包的定義の 2 種類がある。内包的定義とは集合に含まれる条件を定義したものである。たとえば動物と植物を区別する条件を列挙したのが内包的定義である。それに対して、動物＝｛　　｝の中に具体的な動物名、たとえばサンゴと書き込むのが外包的な定義である。

4 マハラノビスの距離への関心は 1930 年代の頭蓋骨の計測データを用いた人類学研究に端を発している。遺跡から発掘された頭蓋骨の計測値をもとに、種族を識別する研究が行われた。当然ながら集落内の個人差よりも集落間の相違が研究の関心事になった。

5 ２つの x, y の関係を示す指標 d が次の公理を満たすとき d を距離と呼ぶ。

(1) $d(x,x)=0,$ かつ $d(x,y)=0$ ならば　$x=y$

(2) $d(x,y)=d(y,x)$

(3) $d(x,y)+d(y,z)\geq d(x,z)$

(2)は対称性、(3)は 3 角不等式と呼ばれる。(1)~(3)の距離の公理から $d(x,y)\geq 0$ という距離の非負性が導かれる。

[6] 一次従属とは、すべてがゼロではない係数$a_1, a_2, \cdots, a_p$ があって変数の組に $a_1X_1 + a_2X_2 + \cdots + a_pX_p = 0$ という関係が成り立つことをいう。したがって変数の合計だけが一次従属のすべてではない。

[7] 永田（2005,105 頁）は別の証明で汎距離の不変性を示している。

[8] p 次の行列 S が $rank(\boldsymbol{S}) = p$ のとき S は正則であるという。このとき $|\boldsymbol{S}| \neq 0$ なので $\boldsymbol{S}^{-1}$ が求められる。一方、$rank(\boldsymbol{S}) < p$ のとき S は特異と呼ばれる。このとき $|\boldsymbol{S}| = 0$ であるから逆行列は計算できない。分散共分散行列が正則でない場合にどのような出力が得られるかは個々のプログラムの作りこみ次第である。たとえば R で 1 m関数を使うと多重共線性を起こしている回帰係数の一つが NA になる。何も出力せずにストップするプログラムもある。

[9] 市販の統計パッケージは、$\theta = 0.5$ で固定するか θ を調査データに任せるという仕組みが少なくない。つまり出現率が低いユーザーを正しく予測したければ大量のノンユーザーを調査しろという予測式になっている。調査データとは無関係に事前確率を入力して判別予測できるように計算法を一般化すべきだろう。

[10] (6.10)は Truett ら (1967)が提唱した多重ロジスティック関数と似ている。しかし多重ロジスティック関数は 2 値の Y を線形モデル $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{b}$ で予測するモデルであり、関数 $f(\boldsymbol{x})$ に多変量正規分布を仮定しない。多重ロジスティック関数から θ に関する事前情報を除いたモデルが、よく知られているロジスティック回帰分析である。

[11] 4.1 章で述べたように複数の母集団についても汎距離は定義されている。したがって汎距離は単一母集団に限る尺度ではない。